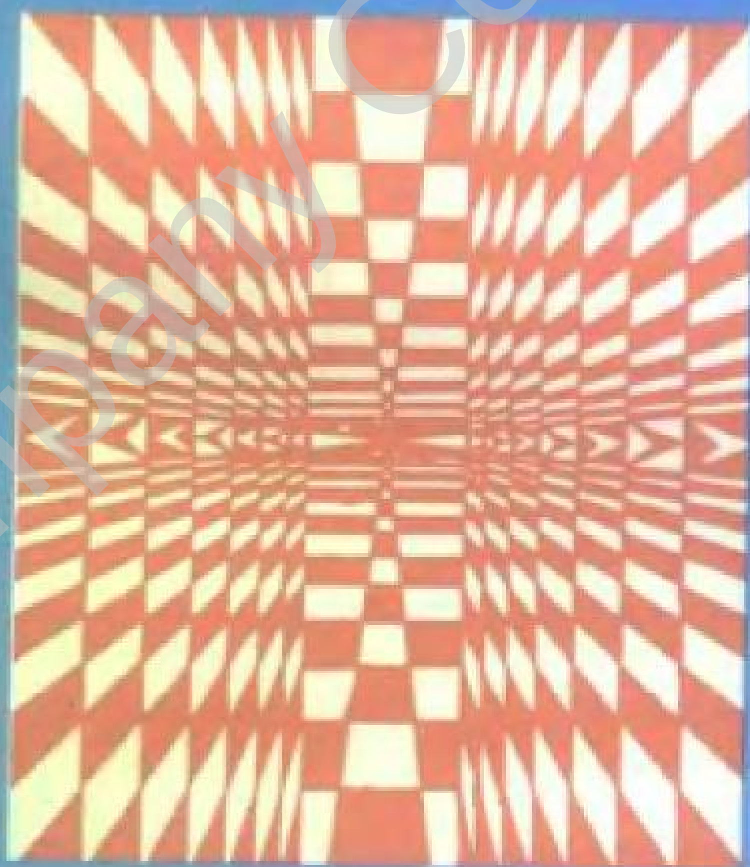
北京大学数学丛书

矩阵计算的理论与方法

徐树方　编著



北京大学出版社

北京大学数学丛书

矩阵计算的理论与方法

徐树方　编著



北京大学出版社
北　　京

新登字(京)159号

图书在版编目(CIP)数据

矩阵计算的理论与方法/徐树方编著.—北京：北京大学出版社，1995.8
(北京大学数学丛书)
ISBN 7-301-02742-7

Ⅰ.矩… Ⅱ.徐… Ⅲ.矩阵-计算方法 Ⅳ.0241.6

书　　　名：矩阵计算的理论与方法
著作责任者：徐树方　编著
责 任 编 辑：刘　勇
标 准 书 号：ISBN 7-301-02742-7/O·[illegible]49
出　版　者：北京大学出版社
地　　　址：北京市海淀区中关村北京大学校内　100871
电　　　话：出版部 2502015　发行部 2559712　编辑部 2502032
排　印　者：北京大学印刷厂
发　行　者：北京大学出版社
经　销　者：新华书店
850×1168毫米　32开本　12印张　302千字
1995年8月第一版　1995年8月第一次印刷
印　　　数：0001—3,000册
定　　　价：19.30元

《北京大学数学丛书》编委会

说　　明

此丛书是以数学、计算数学、概率统计及有关专业的高年级学生、研究生、青年教师及数学研究工作者为读者对象的出版物。丛书特点是内容新颖，力图反映现代数学的新成就；叙述精练，约相当于一学期周学时为3的研究生课程的取材。我们编辑出版此丛书的主要目的是为了适应我们国家培养研究生的需要，同时，又可作为数学及有关系科高年级选修课程的参考书，为提高本科生的教学质量贡献一份力量。

我们诚恳地希望：广大读者对于书目的选择，内容的取材提出宝贵意见，作为我们今后出版或再版时的参考。

《北京大学数学丛书》编委会

一九八一年元月

内容提要

本书系统阐述了矩阵计算这门学科的基础理论、基本方法和近十几年来发展成熟并得到了广泛应用的新成果。内容包括：矩阵知识的复习和补充，矩阵计算概论；求解线性方程组的直接法和迭代法，线性最小二乘问题，共轭梯度法；求解特征值问题的QR方法和同伦方法；Lanczos 方法以及求解 Jacobi 矩阵特征值反问题的正交约化方法等。

本书取材上，既注重基础理论的严谨性、方法的实用性，又保持了内容的新颖性，反映了该学科的最新进展。本书内容自封，各章之间相对独立，可适用于不同读者的需要。

本书可作为计算数学、应用数学等有关专业高年级大学生和研究生的教材或教学参考书，也可供从事科学计算的数学工作者、工程技术人员和高校有关专业的高年级大学生和教师参考。

前　言

写这本书的主要目的，是为计算数学有关专业研究生和高年级大学生提供一本既能反映矩阵计算的基础理论、基本方法和最新进展，又具有实用性和启发性的教学参考书，使之通过这本书的学习，能够对矩阵计算的有关理论和方法有一个比较全面、系统的了解，并为进一步学习与研究，打下一个较好的基础。

基于这样的目的，本书在注重基础理论的前提下，重点放在介绍矩阵计算这门学科近十几年来发展成熟并得到了广泛应用的理论和方法，其中主要包括：不完全分解预优共轭梯度法，广义共轭剩余法，广义极小剩余法，Lanczos 方法，求解特征值问题的同伦方法和分而治之法，以及求解 Jacobi 矩阵特征值反问题的正交约化法。对每一种数值方法，既注重了其数学理论的严谨性，又注重了其实用性，而且还在其来龙去脉上花费了较多的笔墨。每章都精选了几道难易程度不同的习题，用以帮助读者复习、巩固和加深理解有关内容。书末还附有大量较新的参考文献，以便读者进一步深究有关内容时参考。

为了适应各种不同的需要，本书各章内容相对独立，读者可根据自己的需要选择其中任何一部分进行学习，而不会受到前面有关章节内容的影响。其次，为了满足只关心应用的读者的需要，书中对那些较实用的方法都写出了比较详细的算法，根据这些算法去编制适用的软件应该说不会有太大困难。

本书的大部分内容，作者曾在中国科学院研究生院和北京大学为研究生开设的数值代数课上讲授过多次。根据作者的经验，讲授完本书全部内容大约需要80学时左右。此外，本书用到的数学基础，不超过线性代数和数学分析的范畴。因此，一般来说，

对于计算数学和应用数学有关专业的高年级学生和研究生，阅读本书时应该不会感到困难。

作者感谢应隆安教授、滕振寰教授和郭懋正教授的鼓励和支持。计算数学教研室的全体同志曾对本书的编写大纲进行过认真的讨论，并提出许多宝贵的建议和意见，在此向他们表示衷心的感谢。还要感谢孙继广教授，他一直关心着本书的编写和出版，并仔细地审阅了全部手稿。同时还要感谢责任编辑刘勇同志为本书的出版付出的辛勤劳动。

限于水平，书中不当乃至错误之处难免，恳请读者提出批评指正，以期今后改正。

徐树方

1994年9月于北京大学

目 录

第一章　矩阵知识的复习和补充

§1　主要记号和定义

为了以后行文方便和避免不必要的重复，这一节将本书常用的一些记号和术语作一简要的说明。

实数的全体用 $\mathbb{R}$ 表示；实 n 维向量的全体用 $\mathbb{R}^n$ 表示；$m\times n$ 实矩阵的全体用 $\mathbb{R}^{m\times n}$ 表示；对应的复元素的集合分别用 $\mathbb{C}$，$\mathbb{C}^n$ 和 $\mathbb{C}^{m\times n}$ 表示。

对于给定的 $A\in\mathbb{C}^{m\times n}$，我们用 $\bar{A}$，A^{T} 和 A^* 分别表示矩阵 A 的共轭，转置和共轭转置矩阵；用 $\mathrm{rank}(A)$ 表示 A 的秩；用 $\mathbb{C}_r^{m\times n}$ 表示 $\mathbb{C}^{m\times n}$ 中所有秩为 r 的矩阵的全体。

如果 $A\in\mathbb{C}^{n\times n}$，则称 A 为 n 阶**方阵**；对于给定的 n 阶方阵 A，我们用 $\det(A)$ 和 $\mathrm{tr}(A)$ 分别表示 A 的行列式和迹；如果 $\det(A)\neq 0$，就称 A 是**非奇异的**；对于非奇异矩阵 A，用 A^{-1} 表示 A 的逆矩阵。

n 阶单位方阵用 I_n 表示，它的第 k 列用 $e_k^{(n)}$ 表示，即

$$I_n=\begin{bmatrix}1 & & & 0\\ & 1 & & \\ & & \ddots & \\ 0 & & & 1\end{bmatrix}_{n\times n},$$

$$e_k^{(n)}=(0,\cdots,0,\underset{k}{1},0,\cdots,0)^{\mathrm{T}}\in\mathbb{C}^n.$$

当维数从上下文中一目了然时，我们就简记它们为 I 和 e_k。

对于 $A\in\mathbb{C}^{m\times n}$，我们记

$$A = [c_1, \cdots, c_n]$$

是指 $c_k \in \mathbb{C}^m$ 是 A 的第 k 列；同样

$$A = \begin{bmatrix} r_1^{\mathrm{T}} \\ \vdots \\ r_m^{\mathrm{T}} \end{bmatrix}$$

意味着 $r_k \in \mathbb{C}^n$ 是 A 的第 k 行元素构成的 n 维列向量。

设 $\mathscr{X} \subset \mathbb{C}^n$ 是一个子空间。我们用 $\mathscr{X}^{\perp}$ 表示 $\mathscr{X}$ 的正交补空间，即

$$\mathscr{X}^{\perp} = \{x \in \mathbb{C}^n: x^*y = 0, \text{对一切的} y \in \mathscr{X} \text{成立}\}.$$

设 $a_1, \cdots, a_m \in \mathbb{C}^n$。$a_1, \cdots, a_m$ 所有线性组合构成的集合称之为 $a_1, \cdots, a_m$ 张成的**子空间**，记作 $\mathrm{span}\{a_1, \cdots, a_m\}$。

设 $A \in \mathbb{C}^{m \times n}$，有两个与 A 有关的重要的子空间：A 的**值域**（或称**像空间**）

$$\mathscr{R}(A) = \{y \in \mathbb{C}^m: y = Ax, x \in \mathbb{C}^n\},$$

和 A 的**零空间**（或称**核**）

$$\mathscr{N}(A) = \{x \in \mathbb{C}^n: Ax = 0\}.$$

我们用 $\mathrm{diag}(a_1, \cdots, a_n)$ 表示对角元素为 $a_1, \cdots, a_n$ 的 n 阶对角阵，且当 a_i 被方阵代替后，它表示块对角阵。

$\{1, 2, \cdots, n\}$ 的排列的全体用 $\mathscr{S}_n$ 表示；对于任意的 $\pi \in \mathscr{S}_n$，令

$$P_\pi = [e_{\pi(1)}, e_{\pi(2)}, \cdots, e_{\pi(n)}],$$

并称之为对应于 π 的**排列方阵**；n 阶排列方阵的全体记作 $\mathscr{P}_n$。显然 $\mathscr{S}_n$ 和 $\mathscr{P}_n$ 都有 $n!$ 个元素。

设 $A \in \mathbb{C}^{n \times n}$。如果 $A^*A = I$，则称 A 是**酉矩阵**，n 阶酉矩阵的全体记作 $\mathscr{U}_n$；如果 $A^* = A$，则称 A 是 **Hermite 矩阵**；如果 $A^*A = AA^*$，则称 A 是**正规矩阵**。显然，酉矩阵和 Hermite 矩阵都是正规矩阵。实酉矩阵称作**实正交矩阵**；实 Hermite 矩阵称作**实对**

称矩阵，n 阶实对称矩阵的全体记作 $S\mathbb{R}^{n\times n}$。对于一个 Hermite 矩阵 A，如果对任意的非零向量 $x\in\mathbb{C}^n$，都有 $x^*Ax>0$(或$x^*Ax\geq 0$)，则称 A 是**正定的**（或**半正定的**）。

设$A\in\mathbb{C}^{n\times n}$。如果存在 $x\in\mathbb{C}^n$和 $\lambda\in\mathbb{C}$ 满足

$$Ax=\lambda x, \qquad x\neq 0,$$

则称 λ 是 A **的特征值**，x 是 A 属于 λ 的**特征向量**，A 的特征值的全体记作 $\lambda(A)$。容易验证，$\lambda\in\lambda(A)$ 的充分必要条件是

$$\det(\lambda I-A)=0.$$

因此，多项式 $p(\lambda)=\det(\lambda I-A)$ 称作 A 的**特征多项式**。

显然有 $\det(\lambda I-A)=\det(\lambda I-A^{\mathrm{T}})$。因此，$\lambda(A)=\lambda(A^{\mathrm{T}})$。于是，对任意的 $\lambda\in\lambda(A)$,必存在非零向量 $y\in\mathbb{C}^n$，使$A^{\mathrm{T}}y=\lambda y$，即 $y^{\mathrm{T}}A=\lambda y^{\mathrm{T}}$，故称 y 为 A 属于 λ 的**左特征向量**；相对应，属于λ 的特征向量亦称作 A 属于 λ 的**右特征向量**。通常左、右特征向量是不相等的。

如果 A 有 r 个互不相同的特征值 $\lambda_1,\lambda_2,\cdots,\lambda_r$,它们作为$p(\lambda)$的根分别是 $n(\lambda_1),n(\lambda_2),\cdots,n(\lambda_r)$ 重的，即

$$p(\lambda)=\prod_{i=1}^{r}(\lambda-\lambda_i)^{n(\lambda_i)}, \qquad \lambda_i\neq\lambda_j \ \ (i\neq j),$$

$$\sum_{i=1}^{r}n(\lambda_i)=n,$$

则称 $n(\lambda_i)$为 λ_i的**代数重数**；而称

$$m(\lambda_i)=n-\mathrm{rank}(\lambda_i I-A)$$

为λ_i的**几何重数**,它表示 A 之属于 λ_i的线性无关的特征向量的个数。显然有

$$1\leq m(\lambda_i)\leq n(\lambda_i)\leq n.$$

值得指出的是，几何重数和代数重数是两个不同的概念。但习惯上人们将代数重数简称为重数。例如，在今后的叙述中，我们说 λ 是 A 的 p 重特征值，是指 λ 的代数重数是 p 。一般将代数重数为 1 的特征值称作**单特征值**。

如果 $A,B\in\mathbb{C}^{n\times n}$ 满足 $A=X^{-1}BX$，其中 X 是非奇异的 n 阶方阵，则称 A 与 B **相似**。容易验证，如果 A 与 B 相似，则

$$\det(\lambda I-A)=\det(\lambda I-B).$$

因此，相似矩阵有相同的特征多项式，从而它们有相同的特征值。

对于任意的 $A\in\mathbb{C}^{n\times n}$，有下面著名的 Jordan 标准形定理。

定理1.1 设 A 有 r 个互不相同的特征值 $\lambda_1,\cdots,\lambda_r$，其代数重数分别为 $n(\lambda_1),\cdots,n(\lambda_r)$。则必存在非奇异矩阵 $P\in\mathbb{C}^{n\times n}$，使得

$$P^{-1}AP=J\equiv\operatorname{diag}(J(\lambda_1),\cdots,J(\lambda_r)),$$

其中

$$J(\lambda_i)=\operatorname{diag}(J_1(\lambda_i),\cdots,J_{k_i}(\lambda_i))\in\mathbb{C}^{n(\lambda_i)\times n(\lambda_i)},$$

$$1\leqslant i\leqslant r,$$

$$J_k(\lambda_i)=\begin{bmatrix}\lambda_i & 1 & & 0\\ & \lambda_i & \ddots & \\ & & \ddots & 1\\ 0 & & & \lambda_i\end{bmatrix}\in\mathbb{C}^{n_k(\lambda_i)\times n_k(\lambda_i)},\quad 1\leqslant k\leqslant k_i$$

$$\sum_{k=1}^{k_i}n_k(\lambda_i)=n(\lambda_i),\qquad 1\leqslant i\leqslant r;$$

并且除了 $J_k(\lambda_i)$ 的排列次序可以改变外，J 是唯一确定的。

定理1.1中的 J 称作 A 的 **Jordan 标准形**，其中每个 $J_k(\lambda_i)$ 称作 **Jordan 块**。

如果 A 的 Jordan 标准形中每个 Jordan 块都是一阶的，则称

A是**非亏损的**；否则称A是**亏损的**。对于非亏损矩阵有下面重要结果。

定理1.2 设$A\in\mathbb{C}^{n\times n}$。则下面三条等价：

(1) A非亏损；

(2) 存在非奇异矩阵Q使$A=Q\Lambda Q^{-1}$，其中

$$\Lambda=\mathrm{diag}(\lambda_1,\cdots,\lambda_n);$$

(3) $m(\lambda_i)=n(\lambda_i)$，$\lambda_i\in\lambda(A)$。

基于定理1.2的(2)，人们通常亦称非亏损矩阵为**可对角化矩阵**（即可相似于一个对角阵）。

如果A的每个特征值的几何重数都是1，即A属于每个特征值的线性无关的特征向量的个数均为1，则称A是**非减次的**；否则称A是**减次的**。矩阵A非减次等价于A的Jordan标准形中的每个特征值对应的Jordan块是唯一的。

此外，我们用δ_{ij}来表示Kronecker记号，即

$$\delta_{ij}=\begin{cases}0, & i\neq j,\\ 1, & i=j.\end{cases}$$

§2 Schur分解和奇异值分解

这一节，我们介绍两个理论上和应用上都非常重要的矩阵分解定理，即Schur分解定理和奇异值分解定理。

2.1 Schur分解

引理2.1 对于任意的$x=(\alpha_1,\cdots,\alpha_n)^{\mathrm{T}}\in\mathbb{C}^n, x\neq 0$，如果令

$$p=\begin{cases}\|x\|_2, & \text{当 } \alpha_1=0\text{时},\\ -\mathrm{e}^{\mathrm{i}\arg\alpha_1}\|x\|_2, & \text{当 } \alpha_1\neq 0\text{时},\end{cases}\tag{2.1}$$

并定义

$$H(w)=I-2ww^*, \tag{2.2}$$

其中

$$w=(x-pe_1)/\|x-pe_1\|_2, \tag{2.3}$$

则有

$$H(w)x=pe_1. \tag{2.4}$$

证明 直接验证等式（2.4）即可。

容易证明，(2.2)所定义的矩阵 $H(w)$既是 Hermite 矩阵又是酉矩阵，即

$$H(w)^*=H(w)=H(w)^{-1}. \tag{2.5}$$

利用引理2.1和数学归纳法可证下面的定理。

定理2.1（Schur 分解定理） 设 $A\in\mathbb{C}^{n\times n}$. 则存在 $U\in\mathscr{U}_n$，使得

$$U^*AU=T, \tag{2.6}$$

其中 T 是上三角矩阵；而且适当选取 U，可使 T 的对角元素按任意指定的顺序排列。

证明 对矩阵的阶数 n 用数学归纳法。

$n=1$时，定理2.1显然成立。下面考虑 $n>1$的情形，并假定定理的结论对 $n-1$的情形已经成立。

设 $\lambda\in\lambda(A)$是我们希望排在左上角的特征值，并假定 $x\in\mathbb{C}^n$，$x\neq0$，是对应于 λ 的特征向量，即

$$Ax=\lambda x.$$

由引理2.1知，存在 Hermite 酉矩阵 H 和非零常数 p，使得

$$Hx=pe_1.$$

于是

$$HAHe_1=HA\left(\frac{1}{p}x\right)=\lambda e_1.$$

这表明 HAH 有如下形状

$$HAH=\begin{bmatrix}\lambda & * \\ 0 & A_1\end{bmatrix},$$

其中 A_1 是 $n-1$阶方阵。注意到 $\lambda(A_1)=\lambda(A)-\{\lambda\}$，由归纳法假设知，存在 $U_1\in\mathscr{U}_{n-1}$使得

$$U_1^*A_1U_1=T_1,$$

其中 T_1 是上三角矩阵，而且 T_1 的对角元素按我们指定的次序排列。现令

$$U=H\mathrm{diag}(1,U_1),$$

则 $U\in\mathscr{U}_n$，且 U^*AU 具有定理2.1要求的形式。证毕。

注2.1 分解式（2.6）称作 A 的 **Schur 分解**，其右端的 T 称作 A的 **Schur 上三角标准形**，可记作

$$T=\Lambda+M,\quad \Lambda=\mathrm{diag}(\lambda_1,\cdots,\lambda_n),$$

其中M为严格的上三角矩阵，λ_i 为A的特征值。

从定理2.1立即可得

推论2.1 设 $A\in\mathbb{C}^{n\times n}$。则

(1) A是正规矩阵的充分必要条件是存在$U\in\mathscr{U}_n$使得

$$U^*AU=\mathrm{diag}(\lambda_1,\cdots,\lambda_n);$$

(2) A是 Hermite 矩阵的充分必要条件是存在$U\in\mathscr{U}_n$和实对角矩阵 Λ 使得

$$U^*AU=\Lambda.$$

2.2 奇异值分解

定义2.1 设 $A\in\mathbb{C}^{m\times n}$。$A^*A$ 的特征值的非负平方根称作 A 的**奇异值**；A的奇异值的全体记作 $\sigma(A)$。

定理2.2（奇异值分解定理） 设 $A\in\mathbb{C}_r^{m\times n}$。则存在 $U\in\mathscr{U}_m$ 和 $V\in\mathscr{U}_n$，使得

$$U^*AV = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \begin{matrix} r \\ m-r \end{matrix}, \tag{2.7}$$
$$\qquad\qquad r \quad n-r$$

其中 $\Sigma_r = \mathrm{diag}(\sigma_1, \cdots, \sigma_r)$，$\sigma_1 \geqslant \cdots \geqslant \sigma_r > 0$。

证明 由于 A^*A 是半正定的 Hermite 矩阵，且 $\mathrm{rank}(A^*A) = \mathrm{rank}(A)$，故 A^*A 有如下形式的 Schur 分解

$$V^*(A^*A)V = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_n^2), \tag{2.8}$$

其中 $V = [v_1, \cdots, v_n] \in \mathscr{U}_n$，$\sigma_1 \geqslant \cdots \geqslant \sigma_r > 0$，$\sigma_{r+1} = \cdots = \sigma_n = 0$。令

$$V_1 = [v_1, \cdots, v_r], V_2 = [v_{r+1}, \cdots, v_n], \Sigma_r = \mathrm{diag}(\sigma_1, \cdots, \sigma_r).$$

则从（2.8）可得

$$V_1^*(A^*A)V_1 = \Sigma_r^2, \tag{2.9}$$

$$V_2^*(A^*A)V_2 = 0. \tag{2.10}$$

（2.9）表明 AV_1 的列是相互正交的；而(2.10)表明 AV_2 的列都是零向量，即 $AV_2 = 0$。因此，令

$$U_1 = AV_1\Sigma_r^{-1},$$

则 $U_1^*U_1 = I_r$。再取 $U_2 \in \mathbb{C}^{m\times(m-r)}$ 使 $U = [U_1, U_2] \in \mathscr{U}_m$（$\mathbb{C}^m$中的任何一组正交向量都可扩展成全空间的一组正交基），则有

$$U^*AV = \begin{bmatrix} U_1^*AV_1 & U_1^*AV_2 \\ U_2^*AV_1 & U_2^*AV_2 \end{bmatrix} = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix}.$$

注2.2 分解式（2.7）称作 A 的**奇异值分解**，通常简称为 SVD；V 的第 i 列 $v_i = Ve_i$ 称作 A 属于 σ_i 的一个单位**右奇异向量**，U 的第 i 列 $u_i = Ue_i$ 称作 A 属于 σ_i 的一个单位**左奇异向量**。

这里值得指出的是，分解式（2.7）中的 Σ_r 是由 A 唯一确定的，但对每个奇异值 σ_i 对应的单位奇异向量一般则是不唯一的，仅当 $\sigma_i^2(\sigma_i \neq 0)$ 是 A^*A 的单特征值时才唯一(除相差一个单位复

数因子外）；当 σ_i^2 是 A^*A 的重特征值时，对应的单位奇异向量可取作相应特征子空间中的任何一个单位向量。但如果一旦右奇异向量 v_i 已经选定，则相应的左奇异向量 u_i 亦随之而确定下来，即

$$u_i = \sigma_i^{-1} A v_i;$$

反过来，如果一旦 u_i 已选定，则 v_i 亦由

$$A^* u_i = \sigma_i v_i \quad (\sigma_i \neq 0)$$

唯一确定。

注2.3 从定理2.2的证明过程可以看出，当 $A \in \mathbb{R}^{m\times n}$ 时，分解式（2.7）中的 U 和 V 可取作实正交矩阵。

从 A 的奇异值分解，我们可以得到 A 的一些非常有用的信息，下述推论就列举了其中几条最基本的结论。

推论2.2 设 $A \in \mathbb{C}_r^{m\times n}$，则

(1) A 的非零奇异值的个数就等于 $r = \mathrm{rank}(A)$；

(2) $v_{r+1}, \cdots, v_n$ 是 $\mathscr{N}(A)$ 的一组标准正交基；

(3) $u_1, \cdots, u_r$ 是 $\mathscr{R}(A)$ 的一组标准正交基；

(4) $A = \sum\limits_{i=1}^{r} \sigma_i u_i v_i^* = U_1 \Sigma_r V_1^*$（称作 A 的**满秩奇异值分解**）。

作为本节的结束，我们来考察一下奇异值的几何意义。先考虑一个简单的例子。设

$$A = [u_1, u_2] \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} [v_1, v_2]^{\mathrm{T}},$$

其中 $u_1 = (1/2, \sqrt{3}/2)^{\mathrm{T}}, u_2 = (\sqrt{3}/2, -1/2)^{\mathrm{T}}, v_1 = (\sqrt{2}/2, \sqrt{2}/2)^{\mathrm{T}}$，$v_2 = (\sqrt{2}/2, -\sqrt{2}/2)^{\mathrm{T}}$。那么对任意的 $x = \xi_1 v_1 + \xi_2 v_2 \in \mathbb{R}^2$，有

$$y = Ax = \eta_1 u_1 + \eta_2 u_2,$$

其中 $\eta_1 = 3\xi_1$，$\eta_2 = \xi_2$。因此，如果 $\|x\|_2 = 1$，即 $\xi_1^2 + \xi_2^2 = 1$，则对应的 $y = \eta_1 u_1 + \eta_2 u_2$ 满足

$$\frac{\eta_1^2}{3^2}+\eta_2^2=1.$$

这表明A将$\mathbb{R}^2$中的单位圆$\{x\in\mathbb{R}^2: \|x\|_2=1\}$变成了椭圆$E_2=\{y=Ax: \|x\|_2=1\}$，而两个奇异值正好是这一椭圆的两个半轴长；长轴所在的直线是$\mathrm{span}\{u_1\}$，短轴所在的直线是$\mathrm{span}\{u_2\}$，它们分别是$\mathrm{span}\{v_1\}$和$\mathrm{span}\{v_2\}$的像（如图2.1所示）。

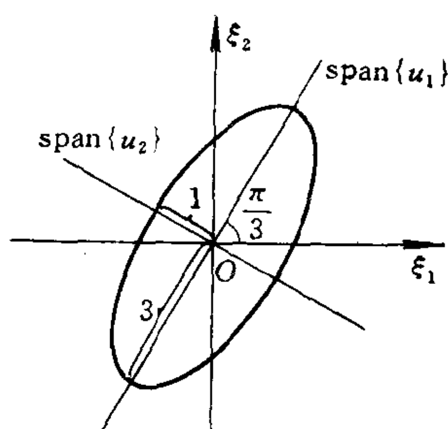


图 2.1

对于一般的情形，不妨设$m=n$，我们亦有

$$E_n=\{y\in\mathbb{C}^n: y=Ax, x\in\mathbb{C}^n, \|x\|_2=1\}$$

是一个超椭球面，它的n个半轴长正好是A的n个奇异值$\sigma_1\geqslant\sigma_2\geqslant\cdots\geqslant\sigma_n\geqslant0$，这些轴所在的直线正好是$A$的左奇异向量所在的直线，它们分别是对应的右奇异向量所在直线的像。

§3 向量范数和矩阵范数

3.1 向量范数

定义3.1 如果定义在$\mathbb{C}^n$上的一个非负实值函数$\|\cdot\|$，对任意的$x, y\in\mathbb{C}^n$和$\alpha\in\mathbb{C}$都有：

(1) 正定性：$x\neq 0\Rightarrow\|x\|>0$,

(2) 齐次性：$\|\alpha x\|=|\alpha|\,\|x\|$,

(3) 半可加性：$\|x+y\|\leqslant\|x\|+\|y\|$,

则称$\|\cdot\|$为 $\mathbb{C}^n$ 上的一个向量范数。

$\mathbb{C}^n$ 上最著名的范数是 p 范数（亦称 Hölder 范数）：

$$\|x\|_p=\left(\sum_{i=1}^{n}|\xi_i|^p\right)^{1/p},\qquad x=(\xi_1,\cdots,\xi_n)^{\mathrm{T}}\in\mathbb{C}^n,\tag{3.1}$$

这里$1\leqslant p\leqslant\infty$。其中最常用的是 $p=1,2,+\infty$时的 p 范数，即

$$\|x\|_1=\sum_{i=1}^{n}|\xi_i|,$$

$$\|x\|_2=\left(\sum_{i=1}^{n}|\xi_i|^2\right)^{1/2},$$

$$\|x\|_\infty=\max_{1\leqslant i\leqslant n}|\xi_i|.$$

关于$\|\cdot\|_1$和$\|\cdot\|_\infty$满足向量范数定义的三条是容易验证的；而要验证$\|\cdot\|_p(1<p<\infty)$是 $\mathbb{C}^n$ 上的向量范数，则需用到著名的 Hölder 不等式：

$$\sum_{i=1}^{n}|\xi_i\eta_i|\leqslant\|x\|_p\|x\|_q,$$

其中 $x=(\xi_1,\cdots,\xi_n)^{\mathrm{T}}$ 和 $y=(\eta_1,\cdots,\eta_n)^{\mathrm{T}}\in\mathbb{C}^n$，$p,q$ 均为大于 1 的实数，且满足$\dfrac{1}{p}+\dfrac{1}{q}=1$。此外，$p$ 范数还有下面的重要性质：

(1) 对任意的 $1\leqslant p<q$ 有

$$\|x\|_q\leqslant\|x\|_p,\qquad\forall x\in\mathbb{C}^n;$$

(2) $\lim\limits_{p\to\infty}\|x\|_p=\|x\|_\infty,\qquad\forall x\in\mathbb{C}^n$.

有关这些结果的证明，这里不再给出，有兴趣的读者可参阅泛函分析方面的有关书籍。

在实际应用中，常常需要利用已知范数去构造出实用的新范数，下述定理给出一种最简单的构造方法。

定理3.1 设$\|\cdot\|$是$\mathbb{C}^m$上的范数，$A\in\mathbb{C}_n^{m\times n}(m\geqslant n)$。则由

$$v_A(x)=\|Ax\|,\ x\in\mathbb{C}^n, \tag{3.2}$$

所定义的非负实函数v_A是$\mathbb{C}^n$上的范数。

证明 直接验证这样定义的v_A满足定义3.1的三条即可。

推论3.1 设$A\in\mathbb{C}^{n\times n}$为正定矩阵。则由

$$\|x\|_A=\sqrt{x^*Ax}\ ,\ x\in\mathbb{C}^n \tag{3.3}$$

定义的$\|\cdot\|_A$是$\mathbb{C}^n$上的范数。

(3.3) 所定义的范数是非常重要的，在某些实际应用中常常是十分方便的。

$\mathbb{C}^n$上的范数的一个非常重要的性质就是范数等价性，即

定理3.2 设$\|\cdot\|$和$\|\cdot\|_*$是$\mathbb{C}^n$上的任意两个范数。则存在正数δ_1和δ_2，使得对任意的$x\in\mathbb{C}^n$都有

$$\delta_1\|x\|\leqslant\|x\|_*\leqslant\delta_2\|x\|, \tag{3.4}$$

即这两个范数是等价的。

证明 首先，对任意的$x=(\xi_1,\cdots,\xi_n)^T\in\mathbb{C}^n$，利用范数的性质和Cauchy不等式，可得

$$\|x\|=\left\|\sum_{i=1}^n\xi_ie_i\right\|\leqslant\sum_{i=1}^n|\xi_i|\,\|e_i\|\leqslant\gamma_2\|x\|_2, \tag{3.5}$$

其中$\gamma_2=\left(\sum_{i=1}^n\|e_i\|^2\right)^{1/2}>0$，与$x$无关。

其次，令

$$\gamma_1=\inf_{y\in\mathbb{S}_{n-1}}\|y\|,$$

其中$\mathbb{S}_{n-1}=\{y\in\mathbb{C}^n:\ \|y\|_2=1\}$。由下确界的定义知，必存在$\{y_k\}_{k=1}^{\infty}\subset\mathbb{S}_{n-1}$，使得

$$\lim_{k\to\infty}\|y_k\|=\gamma_1.$$

再由 Weierstrass 聚点原则和 S_{n-1} 的有界闭性知，$\{y_k\}_{k=1}^{\infty}$ 必含有一个收敛于 S_{n-1} 中某点的子序列，现不妨就假定 $\{y_k\}_{k=1}^{\infty}$ 收敛于 $y_0 \in S_{n-1}$，即

$$\lim_{k\to\infty} \|y_k - y_0\|_2 = 0.$$

由范数的半可加性和 (3.5)，可得

$$|\|y_k\| - \|y_0\|| \leqslant \|y_k - y_0\| \leqslant \gamma_2 \|y_k - y_0\|_2,$$

从而有

$$\gamma_1 = \lim_{k\to\infty} \|y_k\| = \|y_0\| > 0.$$

现对任意的非零向量 $x \in C^n$，有 $y = x/\|x\|_2 \in S_{n-1}$，因此

$$\|x\| = \|(\|x\|_2)y\| = \|x\|_2 \|y\| \geqslant \gamma_1 \|x\|_2. \tag{3.6}$$

由(3.5)和(3.6)给出

$$\gamma_1 \|x\|_2 \leqslant \|x\| \leqslant \gamma_2 \|x\|_2 \tag{3.7}$$

对一切的 $x \in C^n$ 成立，其中 γ_1 和 γ_2 与 x 无关。

同理可找到与 x 无关的正数 η_1 和 η_2，使得

$$\eta_1 \|x\|_2 \leqslant \|x\|_* \leqslant \eta_2 \|x\|_2 \tag{3.8}$$

对一切的 $x \in C^n$ 成立。

由(3.7)和(3.8)即知(3.4)成立，其中 $\delta_1 = \eta_1/\gamma_2$，$\delta_2 = \eta_2/\gamma_1$ 是与 x 无关的正数。证毕。

对于常用的三个 p 范数，容易验证，对任意的 $x \in C^n$，有

$$\|x\|_\infty \leqslant \|x\|_1 \leqslant n\|x\|_\infty,$$

$$\frac{1}{\sqrt{n}}\|x\|_1 \leqslant \|x\|_2 \leqslant \|x\|_1,$$

$$\frac{1}{\sqrt{n}}\|x\|_2 \leqslant \|x\|_\infty \leqslant \|x\|_2.$$

用拓扑学的语言来讲，定理3.2是说，尽管在 $\mathbb{C}^n$ 上可有各种各样的范数，但它们所诱导出的拓扑却是一样的，即它们有相同的开集和闭集。因此，在 $\mathbb{C}^n$ 上关于各种范数的连续性概念是一样的。例如，下述定理指出了极限定义的等价性.

定理3.3 设 $\|\cdot\|$ 是 $\mathbb{C}^n$ 上的任一范数，$x_k=(\xi_1^{(k)},\cdots,\xi_n^{(k)})^T\in\mathbb{C}^n$，$k=0,1,2,\cdots$。则 $\lim\limits_{k\to\infty}\|x_k-x_0\|=0$ 当且仅当 $\lim\limits_{k\to\infty}\xi_i^{(k)}=\xi_i^{(0)}$，$i=1,2,\cdots,n$，即范数收敛等价于坐标收敛。

证明留作练习。

3.2 矩阵范数

定义3.2 如果定义在 $\mathbb{C}^{n\times n}$ 上的一个非负实值函数 $\|\cdot\|$，对任意的 $A,B\in\mathbb{C}^{n\times n}$ 和 $a\in\mathbb{C}$ 都有

(1) 正定性：$A\neq 0\Rightarrow\|A\|>0$，

(2) 齐次性：$\|aA\|=|a|\|A\|$，

(3) 半可加性：$\|A+B\|\leqslant\|A\|+\|B\|$，

(4) 相容性：$\|AB\|\leqslant\|A\|\|B\|$，

则称 $\|\cdot\|$ 为 $\mathbb{C}^{n\times n}$ 上的**矩阵范数**。

$\mathbb{C}^{n\times n}$ 上一个常用而又易于定义的矩阵范数就是

$$\|A\|_F=\left(\sum_{i,j=1}^{n}|a_{ij}|^2\right)^{1/2},\qquad A=[a_{ij}]\in\mathbb{C}^{n\times n}. \tag{3.9}$$

通常称作 Frobenius 范数，有时亦称作 Euclid 范数，因它是 $\mathbb{C}^n$ 上的 Euclid 范数 $\|\cdot\|_2$ 的自然推广。此外，利用奇异值分解定理，容易证明

$$\|A\|_F=\left(\sum_{i=1}^{n}\sigma_i^2\right)^{1/2}, \tag{3.10}$$

其中 $\sigma_1,\sigma_2,\cdots,\sigma_n$ 是 A 的奇异值。

注3.1 由于我们可以将 $\mathbb{C}^{n\times n}$ 中的矩阵 A 看作 $\mathbb{C}^{n^2}$ 中的一个向量，因而 $\mathbb{C}^{n\times n}$ 上的任一矩阵范数都可以看作 $\mathbb{C}^{n^2}$ 上的一个

向量范数。于是，根据定理3.2和3.3可得：

（1）$\mathbb{C}^{n\times n}$ 上的任何两个范数都是等价的；

（2）范数收敛等价于矩阵元素收敛，即

$$\lim_{k\to\infty}\|A^{(k)}-A^{(0)}\|=0\Leftrightarrow\lim_{k\to\infty}\max_{i,j}|a_{ij}^{(k)}-a_{ij}^{(0)}|=0,$$

其中 $A^{(k)}=[a_{ij}^{(k)}]\in\mathbb{C}^{n\times n}$，$k=0,1,2,\cdots$，$\|\cdot\|$是 $\mathbb{C}^{n\times n}$ 上的任一矩阵范数。

注3.2 由于向量序列和矩阵序列在各种范数下的收敛是等价的，因此以后在谈到它们的收敛时，就不再特别声明在何种范数意义下的收敛。例如，若$\{A^{(k)}\}_{k=0}^{\infty}\subset\mathbb{C}^{n\times n}$，我们写

$$\lim_{k\to\infty}A^{(k)}=A^{(0)},$$

应理解为对某种矩阵范数$\|\cdot\|$有

$$\lim_{k\to\infty}\|A^{(k)}-A^{(0)}\|=0.$$

下面我们讨论由向量范数诱导出的矩阵范数，这类矩阵范数是矩阵分析中一类十分重要的范数。

定理3.4 设$\|\cdot\|$是 $\mathbb{C}^n$ 上的一个向量范数。则由

$$|||A|||=\max_{\|x\|=1}\|Ax\|,\qquad A\in\mathbb{C}^{n\times n}\tag{3.11}$$

所定义的实值函数 $|||\cdot|||$ 是 $\mathbb{C}^{n\times n}$ 上的一个矩阵范数。

证明 首先，由范数等价定理知，点集

$$\{x\in\mathbb{C}^n:\ \|x\|=1\}$$

是 $\mathbb{C}^n$ 中的有界闭集，而且$\|\cdot\|$是 $\mathbb{C}^n$ 上的连续函数，因此，由(3.11)所定义的 $|||\cdot|||$ 是有意义的。

其次，从（3.11）易知，对任意的 $x\in\mathbb{C}^n$，$x\neq0$，有

$$\frac{\|Ax\|}{\|x\|}=\left\|A\frac{x}{\|x\|}\right\|\leqslant|||A|||,$$

从而有

$$\|Ax\| \leqslant |||A||| \, \|x\|, \qquad \forall x \in \mathbb{C}^n. \tag{3.12}$$

下面证明$|||\cdot|||$满足定义3.2的四条。

(1) 设$A \in \mathbb{C}^{n\times n}$，$A \neq 0$。则必有非零向量$x \in \mathbb{C}^n$，使$Ax \neq 0$。从而由向量范数的正定性和(3.12)有

$$0 < \|Ax\| \leqslant |||A||| \, \|x\|. \tag{3.13}$$

而$\|x\| > 0$，故有$|||A||| > 0$，即$|||\cdot|||$满足正定性。

(2) 任取$\alpha \in \mathbb{C}$，$A \in \mathbb{C}^{n\times n}$，有

$$|||\alpha A||| = \max_{\|x\|=1} \|\alpha Ax\| = |\alpha| \max_{\|x\|=1} \|Ax\| = |\alpha| \, |||A|||,$$

即$|||\cdot|||$满足齐次性。

(3) 任取$A, B \in \mathbb{C}^{n\times n}$。设$x \in \mathbb{C}^n$满足

$$\|x\| = 1 \text{ 和 } \|(A+B)x\| = |||A+B|||.$$

则由(3.12)和向量范数的半可加性有

$$\begin{aligned} |||A+B||| &= \|(A+B)x\| \leqslant \|Ax\| + \|Bx\| \\ &\leqslant |||A||| \, \|x\| + |||B||| \, \|x\| = |||A||| + |||B|||, \end{aligned}$$

即$|||\cdot|||$满足半可加性。

(4) 任取$A, B \in \mathbb{C}^{n\times n}$。设$x \in \mathbb{C}^n$满足

$$\|x\| = 1 \text{ 和 } \|ABx\| = |||AB|||.$$

则由(3.12)有

$$\begin{aligned} |||AB||| &= \|ABx\| \leqslant |||A||| \, \|Bx\| \\ &\leqslant |||A||| \, |||B||| \, \|x\| = |||A||| \, |||B|||, \end{aligned}$$

即$|||\cdot|||$满足相容性。

定义3.3 设$\|\cdot\|$是$\mathbb{C}^n$上的一个向量范数。由(3.11)所定义的矩阵范数$|||\cdot|||$称作$\mathbb{C}^{n\times n}$上由$\|\cdot\|$诱导出的**算子范数**（简

称算子范数），有时亦称作从属于向量范数$\|\cdot\|$的**矩阵范数**。

对 $\mathbb{C}^n$ 上的 p 范数，根据定理3.4，可以得到 $\mathbb{C}^{n\times n}$ 上的算子范数 $|||\cdot|||_p$:

$$|||A|||_p=\max_{\|x\|_p=1}\|Ax\|_p,\qquad A\in\mathbb{C}^{n\times n},\tag{3.14}$$

此处$1\leqslant p\leqslant\infty$。关于 $p=1,2,\infty$所对应的$|||\cdot|||_p$，有如下的表示定理。

定理3.5 设 $A=[a_{ij}]\in\mathbb{C}^{n\times n}$。则有

(1) $|||A|||_1=\max\limits_{1\leqslant j\leqslant n}\sum\limits_{i=1}^{n}|a_{ij}|$ （列和范数），

(2) $|||A|||_\infty=\max\limits_{1\leqslant i\leqslant n}\sum\limits_{j=1}^{n}|a_{ij}|$ （行和范数），

(3) $|||A|||_2=\sigma_1$ （谱范数），

其中 σ_1 表示A的最大奇异值。

这一定理的证明留作练习请读者自己给出。

由定理 3.5 可知，矩阵的列和范数和行和范数很容易由矩阵的元素直接算出；但矩阵的谱范数需要求A的最大奇异值，而这并非易事，因而A的谱范数很少用于实际计算。可是，由于谱范数有许多良好的性质，因此在理论研究中使用起来特别方便。下述定理列举了谱范数的几条常用的性质。

定理3.6 设 $A\in\mathbb{C}^{n\times n}$。则

(1) $|||A|||_2=\max\limits_{\|x\|_2=\|y\|_2=1}|y^*Ax|$,

(2) $|||A^*|||_2=|||A^{\mathrm{T}}|||_2=|||A|||_2$,

(3) $|||A^*A|||_2=|||A|||_2^2$,

(4) $|||A|||_2^2\leqslant|||A|||_1|||A|||_\infty$,

(5) $|||UAV|||_2=|||A|||_2$, $\forall U,V\in\mathscr{U}_n$.

证明留作练习。

注3.3 上面讨论中，为了不致引起混淆，我们将算子范数

记作$|||\cdot|||$。而以后为了书写简单起见，我们亦将算子范数简记为$\|\cdot\|$；特别，对于p范数诱导出的算子范数$|||\cdot|||_p$，简记为$\|\cdot\|_p$。

注3.4 这里为了叙述上的简便，仅对方阵的范数进行了讨论，而所得到的大部分结论都可照搬到非方阵的情形（此时，在矩阵范数的定义中就不能再要求满足(iv)）。例如(3.14)所定义的算子范数，对$A\in\mathbb{C}^{m\times n}$亦成立，且相应的定理3.5亦真，并有

$$\|AB\|_p\leqslant\|A\|_p\|B\|_p,\qquad 1\leqslant p\leqslant\infty,$$

对任意的$A\in\mathbb{C}^{m\times n}$和$B\in\mathbb{C}^{n\times q}$成立。

3.3 谱半径和矩阵序列的收敛性

定义3.4 设$A\in\mathbb{C}^{n\times n}$. 则称

$$\rho(A)=\max\{|\lambda|:\ \lambda\in\lambda(A)\}\tag{3.15}$$

为A的谱半径。

谱半径和矩阵范数之间有如下关系。

定理3.7 设$A\in\mathbb{C}^{n\times n}$。则有：

（1）对$\mathbb{C}^{n\times n}$上的任一矩阵范数$\|\cdot\|$，有

$$\rho(A)\leqslant\|A\|;$$

（2）对于任给的$\varepsilon>0$，存在$\mathbb{C}^{n\times n}$上的算子范数$\|\cdot\|$，使得

$$\|A\|\leqslant\rho(A)+\varepsilon.$$

证明 （1）设$x\in\mathbb{C}^n$满足

$$x\neq 0,\ \ Ax=\lambda x,\ \ |\lambda|=\rho(A),$$

则有

$$\rho(A)\|xe_1^{\mathrm T}\|=\|\lambda xe_1^{\mathrm T}\|=\|Axe_1^{\mathrm T}\|\leqslant\|A\|\|xe_1^{\mathrm T}\|,$$

从而有

$$\rho(A)\leqslant\|A\|.$$

（2）由 Schur 分解定理知，存在 $U\in\mathscr{U}_n$，使得

$$U^*AU=\Lambda+T,$$

其中 $\Lambda=\mathrm{diag}(\lambda_1,\cdots,\lambda_n)$，$T=[t_{ij}]$是严格的上三角矩阵．对于任给 $\delta>0$，令

$$D_\delta=\mathrm{diag}\ (1,\delta,\delta^2,\cdots,\delta^{n-1}),$$

则

$$D_\delta^{-1}U^*AUD_\delta=\Lambda+D_\delta^{-1}TD_\delta$$

$$=\Lambda+\begin{bmatrix} 0 & \delta t_{12} & \delta^2 t_{13} & \cdots & \delta^{n-1}t_{1n} \\ 0 & 0 & \delta t_{23} & \cdots & \delta^{n-2}t_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \delta t_{n-1,n} \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

现对给定的 $\varepsilon>0$，取定 $\delta>0$，使

$$\sum_{i=1}^{j-1}|\delta^{j-i}t_{ij}|\leqslant\varepsilon,\qquad j=2,3,\cdots,n,$$

并定义

$$\|G\|=\|D_\delta^{-1}U^*GUD_\delta\|_1,\qquad \forall G\in\mathbb{C}^{n\times n}.$$

则易证这样定义的函数$\|\cdot\|$是如下定义的向量范数

$$\|x\|_{UD_\delta}=\|(UD_\delta)^{-1}x\|_1,\qquad x\in\mathbb{C}^n$$

诱导出的算子范数，且有

$$\|A\|=\|D_\delta^{-1}U^*AUD_\delta\|_1\leqslant\|\Lambda\|_1+\|D_\delta^{-1}TD_\delta\|_1\leqslant\rho(A)+\varepsilon.$$

定理3.8 设 $A\in\mathbb{C}^{n\times n}$．则

$$\lim_{k\to\infty}A^k=0\Longleftrightarrow\rho(A)<1.$$

证明 *必要性* 设$\lim\limits_{k\to\infty}A^k=0$，并假定$\lambda\in\lambda(A)$满足$\rho(A)$

$=|\lambda|$。由于对任意的 k 有 $\lambda^k \in \lambda(A^k)$，故由定理 3.7 有

$$\rho(A)^k = |\lambda|^k \leqslant \rho(A^k) \leqslant \|A^k\|_2$$

对一切的 k 成立，从而 $\lim\limits_{k\to\infty} \rho(A)^k = 0$，因此必有 $\rho(A)<1$。

充分性 设 $\rho(A)<1$，则由定理 3.7 知，必有算子范数 $\|\cdot\|$，使得 $\|A\|<1$，从而

$$0 \leqslant \|A^k\| \leqslant \|A\|^k \to 0, \qquad k\to\infty,$$

于是 $\lim\limits_{k\to\infty} A^k = 0$。证毕。

利用定理 3.8 容易证明下述重要的定理。

定理3.9 设 $A\in \mathbf{C}^{n\times n}$。则有：

(1) $\sum\limits_{k=0}^{\infty} A^k$ 收敛的充分必要条件是 $\rho(A)<1$；

(2) 当 $\sum\limits_{k=0}^{\infty} A^k$ 收敛时，有

$$\sum_{k=0}^{\infty} A^k = (I-A)^{-1},$$

而且存在 $\mathbf{C}^{n\times n}$ 上的算子范数 $\|\cdot\|$，使得

$$\left\|(I-A)^{-1} - \sum_{k=0}^{m} A^k\right\| \leqslant \frac{\|A\|^{m+1}}{1-\|A\|},$$

对一切的自然数 m 成立。

证明留作练习。

定理3.10 设 $A\in \mathbf{C}^{n\times n}$，$\|\cdot\|$是 $\mathbf{C}^{n\times n}$上的任一矩阵范数。则

$$\lim_{k\to\infty} \|A^k\|^{1/k} = \rho(A)。$$

证明 首先由 $\rho(A^k)\leqslant \|A^k\|$和 $\rho(A^k)=\rho(A)^k$ 可得

$$\rho(A) \leqslant \|A^k\|^{1/k}, \qquad k=1,2,\cdots. \tag{3.16}$$

其次，对任给的 $\varepsilon>0$，令

$$B_\varepsilon=\frac{1}{\rho(A)+\varepsilon}A,$$

则 $\rho(B_\varepsilon)<1$，于是有 $\lim\limits_{k\to\infty}B_\varepsilon^k=0$。因此，必存在 k_0，使当 $k>k_0$ 时，有

$$\|B_\varepsilon^k\|<1,$$

即

$$\|A^k\|<(\rho(A)+\varepsilon)^k. \tag{3.17}$$

由(3.16)和(3.17)知，对任给的 $\varepsilon>0$，已找到了 k_0，使当 $k>k_0$ 时，有

$$\rho(A)\leqslant\|A^k\|^{1/k}\leqslant\rho(A)+\varepsilon,$$

从而

$$\lim_{k\to\infty}\|A^k\|^{1/k}=\rho(A).$$

§4 正交投影和子空间之间的距离

4.1 正交投影

定义4.1 设 $\mathscr{X}\subset\mathbb{C}^n$ 是一个子空间．如果 $P\in\mathbb{C}^{n\times n}$ 满足：

(1) $\mathscr{R}(P)=\mathscr{X}$,

(2) $P^2=P$,

(3) $P^*=P$,

则称 P 是映射到 $\mathscr{X}$ 上的**正交投影**。

从定义可知：正交投影是一个 Hermite 幂等矩阵；对任意的 $x\in\mathbb{C}^n$，有 $Px\in\mathscr{X}$，$(I-P)x\in\mathscr{X}^\perp$；$(I-P)$是映射到 $\mathscr{X}^\perp$ 上的正交投影。

例4.1 设 $v\in\mathbb{C}^n$，$\|v\|_2=1$．则 $P=vv^*$ 是映射到 $\mathscr{X}=\mathrm{span}\{v\}$ 上的正交投影，其几何意义如图4.1所示。

定理4.1 对于任意的子空间 $\mathscr{X}\subset\mathbb{C}^n$，在 $\mathbb{C}^{n\times n}$ 中有且仅有

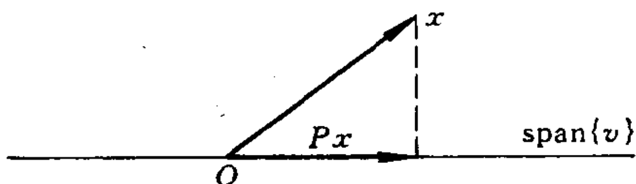


图 4.1

一个映射到$\mathscr{X}$上的正交投影.

证明 *存在性* 设$\mathscr{X}=\mathscr{R}(V)$，其中$V\in\mathbb{C}^{n\times l}$满足$V^*V=I$，即$V$的列构成$\mathscr{X}$的一组标准正交基. 令

$$P=VV^*, \tag{4.1}$$

则容易验证P是映射到$\mathscr{X}$上的正交投影.

唯一性 设P_1也是映射到$\mathscr{X}$上的正交投影. 则对任意的$x\in\mathbb{C}^n$，有

$$\|Px-P_1x\|_2^2=(Px)^*(I-P_1)x+(P_1x)^*(I-P)x=0,$$

从而$P=P_1$. 证毕.

鉴于定理4.1，对于给定的子空间$\mathscr{X}$，通常用$P_{\mathscr{X}}$来表示映射到它上的唯一的正交投影. 此外，尽管对一个子空间来说，它的标准正交基是不唯一的，但按(4.1)定义的矩阵却是唯一的，与正交基的选取无关.

4.2 子空间之间的距离

借助于子空间和映射到它上的正交投影之间的一一对应关系，我们能够给出子空间之间的距离的概念.

假设$\mathscr{X}$和$\mathscr{Y}$是$\mathbb{C}^n$中的两个同维数的子空间. 我们定义这两个子空间之间的距离为

$$\operatorname{dist}(\mathscr{X},\mathscr{Y})=\|P_{\mathscr{X}}-P_{\mathscr{Y}}\|_2, \tag{4.2}$$

其中$P_{\mathscr{X}}$和$P_{\mathscr{Y}}$分别是映射到$\mathscr{X}$和$\mathscr{Y}$上的正交投影.

注4.1 如果我们记$\mathbb{G}_l^n$为$\mathbb{C}^n$中所有l维子空间的全体，则按(4.2)所定义的实值函数$\operatorname{dist}(\cdot,\cdot)$是$\mathbb{G}_l^n$上的距离，即它

满足：

(1) $\text{dist}(\mathscr{Z},\mathscr{W})\geqslant 0$，而且 $\text{dist}(\mathscr{Z},\mathscr{W})=0\Longleftrightarrow \mathscr{Z}=\mathscr{W}$；

(2) $\text{dist}(\mathscr{Z},\mathscr{W})=\text{dist}(\mathscr{W},\mathscr{Z})$；

(3) $\text{dist}(\mathscr{Z},\mathscr{W})\leqslant\text{dist}(\mathscr{Z},\mathscr{X})+\text{dist}(\mathscr{X},\mathscr{W})$，

其中$\mathscr{Z},\mathscr{W}$和$\mathscr{X}$是 G_l^n 中的任意点.

下面我们给出(4.2)所定义的子空间之间的距离的几何解释.

为了简单起见，我们先考虑二维实空间 $\mathbb{R}^2$ 中的两个一维子空间$\mathscr{X}=\text{span}\{x\}$和$\mathscr{Y}=\text{span}\{y\}$（即两条过原点的直线）之间的距离. 此时，假设它们之间的夹角为$\theta(0\leqslant\theta\leqslant\pi/2)$，则通过简单的演算可证

$$\text{dist}(\mathscr{X},\mathscr{Y})=\sin\theta,$$

即按(4.2)所定义的$\mathscr{X}$和$\mathscr{Y}$之间的距离实质上就是这两条直线之间夹角的正弦(参见图4.2)

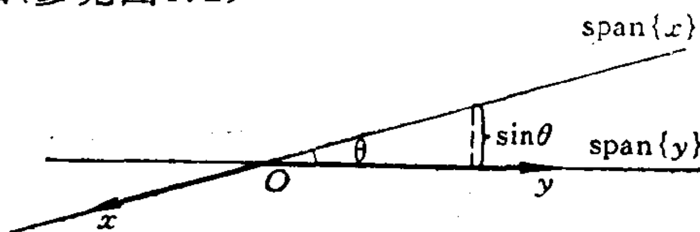


图 4.2

对于一般情形，也有类似的几何解释. 为此，先利用奇异值分解证明一条关于分块酉矩阵的分解定理，即所谓的C-S分解定理.

定理4.2（**C-S 分解定理**） 设酉矩阵 Q 分块为

$$Q=\begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}\begin{matrix} k \\ j \end{matrix}, \quad k\leqslant j.$$
$$\qquad\quad k \qquad j$$

则存在酉矩阵 $U=\text{diag}(\underset{k}{U_1},\underset{j}{U_2})$和 $V=\text{diag}(\underset{k}{V_1},\underset{j}{V_2})$，使得

$$U^*QV=\begin{bmatrix} C & S & 0 \\ -S & C & 0 \\ 0 & 0 & I \end{bmatrix}\begin{matrix} k \\ k \\ j-k \end{matrix}, \qquad (4.3)$$
$$\qquad k \qquad k \qquad j-k$$

其中

$$C=\mathrm{diag}(c_1,\cdots,c_k),\quad c_i=\cos\theta_i,$$
$$S=\mathrm{diag}(s_1,\cdots,s_k),\quad s_i=\sin\theta_i,$$
$$\pi/2\geqslant\theta_1\geqslant\theta_2\geqslant\cdots\geqslant\theta_k\geqslant 0.$$

证明 分四步证之。

(1) 对 Q_{11} 应用奇异值分解定理知，存在 $U_1,V_1\in\mathscr{U}_k$. 使得

$$U_1^*Q_{11}V_1=C,\tag{4.4}$$

其中 $C=\mathrm{diag}(c_1,\cdots,c_k)$，$0\leqslant c_1\leqslant\cdots\leqslant c_k$. 由于 Q 是酉矩阵，故必有 $c_k\leqslant 1$. 为下面的讨论明确起见，现假定 c_i 中有 l 个小于 1，即

$$0\leqslant c_1\leqslant\cdots\leqslant c_l<c_{l+1}=c_{l+2}=\cdots=c_k=1,\quad 0\leqslant l\leqslant k.$$

记 $C_1=\mathrm{diag}(c_1,\cdots,c_l)$.

(2) 存在 $\tilde{U}_2,V_2\in\mathscr{U}_j$，使得

$$\tilde{U}_2^*Q_{21}V_1=\begin{bmatrix}-S\\0\end{bmatrix}\begin{matrix}k\\j-k\end{matrix},\quad U_1^*Q_{12}V_2=\underset{\quad k\qquad j-k}{[S,\ 0]}\,k,$$

其中 $S=\mathrm{diag}(s_1,\cdots,s_k)$，$s_1\geqslant\cdots\geqslant s_k\geqslant 0$，$s_i^2+c_i^2=1$.

从 Q 是酉矩阵的假定和(4.4)，可得

$$I=\left(\begin{bmatrix}Q_{11}\\Q_{21}\end{bmatrix}V_1\right)^*\left(\begin{bmatrix}Q_{11}\\Q_{21}\end{bmatrix}V_1\right)=C^2+V_1^*Q_{21}^*Q_{21}V_1.$$

从而有

$$V_1^*Q_{21}^*Q_{21}V_1=I-C^2=\mathrm{diag}(\underset{l}{I-C_1^2},\ \underset{k-l}{0}).\tag{4.5}$$

上式表明，$Q_{21}V_1$ 的前 l 列互相正交，后 $k-l$ 列为零，即 $Q_{21}V_1$ 有如下形式:

$$Q_{21}V_1=[\underset{l}{W},\ \underset{k-l}{0}],\quad W^*W=I-C_1^2.\tag{4.6}$$

令 $s_i = \sqrt{1-c_i^2}, i=1,2,\cdots,k$，则 $s_1 \geqslant \cdots \geqslant s_l > s_{l+1} = \cdots = s_k = 0$。再令 $S_1 = \operatorname{diag}(s_1,\cdots,s_l)$，$S = \operatorname{diag}(\underset{l}{S_1}, \underset{k-l}{0})$，$\tilde{U}_{21} = -WS_1^{-1}$，则由(4.6)知 $\tilde{U}_{21}^*\tilde{U}_{21} = I$。取 $\tilde{U}_{22} \in \mathbb{C}^{j\times(j-l)}$，使 $\tilde{U}_2 = [\tilde{U}_{21}, \tilde{U}_{22}] \in \mathscr{U}_j$，则有

$$\tilde{U}_2^* Q_{21} V_1 = \begin{bmatrix} \tilde{U}_{21}^* \\ U_{22}^* \end{bmatrix} [W, 0] = \begin{bmatrix} -S_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{matrix} l \\ j-l \end{matrix} \quad = \begin{bmatrix} -S \\ 0 \end{bmatrix} \begin{matrix} k \\ j-k \end{matrix}.$$
$$\qquad\qquad\qquad\qquad l \quad k-l$$

同理可证存在 $V_2 \in \mathscr{U}_j$，使得

$$U_1^* Q_{12} V_2 = [S, 0].$$

(3) 令 $\tilde{U} = \operatorname{diag}(U_1, \tilde{U}_2)$，$V = \operatorname{diag}(V_1, V_2)$，$X = \tilde{U}^* Q V$。则由前面所证知 X 具有如下的分块形式：

$$X = \begin{bmatrix} C_1 & 0 & S_1 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ -S_1 & 0 & X_{33} & X_{34} & X_{35} \\ 0 & 0 & X_{43} & X_{44} & X_{45} \\ 0 & 0 & X_{53} & X_{54} & X_{55} \end{bmatrix} \begin{matrix} \\ k-l \\ l \\ k-l \\ j-k \end{matrix}.$$
$$\qquad l \quad k-l \quad l \quad k-l \quad j-k$$

注意到 X 仍然是酉矩阵，S_1 是对角元素均为正数的对角阵，比较 $X^*X = I$ 的两边可知，$X_{33} = C_1$，X_{34}, X_{35}, X_{43} 和 X_{53} 皆为零矩阵，而且

$$\begin{bmatrix} X_{44} & X_{45} \\ X_{54} & X_{55} \end{bmatrix}$$

是酉矩阵。

(4) 记

$$U_{33} = \begin{bmatrix} X_{44} & X_{45} \\ X_{54} & X_{55} \end{bmatrix},$$

则有

$$\text{diag}(I_{k+l}, U_{33}^*)X = \begin{bmatrix} C_1 & 0 & S_1 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ -S_1 & 0 & C_1 & 0 & 0 \\ 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & I \end{bmatrix}.$$

所以令

$$U_2 = \tilde{U}_2 \text{diag}\ (I_l, U_{33}),$$

则 $\text{diag}(U_1^*, U_2^*)Q\ \text{diag}(V_1, V_2)$具有所要求的形式．证毕．

利用C-S 分解定理可证

定理4.3 设 $\mathscr{X} = \mathscr{R}(X_1)$， $\mathscr{Y} = \mathscr{R}(Y_1)$， 其中 $X_1, Y_1 \in \mathbb{C}^{n \times k}$， $X_1^* X_1 = Y_1^* Y_1 = I$．则

$$\text{dist}(\mathscr{X}, \mathscr{Y}) = (1 - \sigma^2_{\min}(X_1^* Y_1))^{1/2},$$

其中 $\sigma_{\min}(X_1^* Y_1)$表示 $X_1^* Y_1$ 的最小奇异值．

证明 取 $X_2, Y_2 \in \mathbb{C}^{n \times (n-k)}$， 使 $X = [X_1, X_2]$， $Y = [Y_1, Y_2] \in \mathscr{U}_n$．无妨假定 $k \leqslant n-k$， 否则考虑 $\tilde{X} = [X_2, X_1]$和 $\tilde{Y} = [Y_2, Y_1]$即可．对酉矩阵

$$X^* Y = \begin{bmatrix} X_1^* Y_1 & X_1^* Y_2 \\ X_2^* Y_1 & X_2^* Y_2 \end{bmatrix} \begin{matrix} k \\ n-k \end{matrix}$$
$$\qquad\qquad k \qquad\quad n-k$$

应用 C-S 分解定理可知，存在 $U_1, V_1 \in \mathscr{U}_k$ 和 $U_2, V_2 \in \mathscr{U}_{n-k}$，使得

$$U_1^*(X_1^* Y_1)V_1 = C,$$
$$U_1^*(X_1^* Y_2)V_2 = [S, 0],$$
$$U_2^*(X_2^* Y_1)V_1 = \begin{bmatrix} -S \\ 0 \end{bmatrix},$$

其中

$$S = \text{diag}(s_1, \cdots, s_k), \qquad s_i = \sin\theta_i \ \ (i = 1, 2, \cdots, k),$$
$$C = \text{diag}(c_1, \cdots, c_k), \qquad c_i = \cos\theta_i \ \ (i = 1, 2, \cdots, k),$$

$$\pi/2 \geqslant \theta_1 \geqslant \theta_2 \geqslant \cdots \geqslant \theta_k \geqslant 0.$$

于是

$$\begin{aligned}
\mathrm{dist}(\mathscr{X}, \mathscr{Y}) &= \|P_{\mathscr{X}} - P_{\mathscr{Y}}\|_2 = \|X_1X_1^* - Y_1Y_1^*\|_2 \\
&= \|X^*[X_1X_1^* - Y_1Y_1^*]Y\|_2 \\
&= \left\|\begin{bmatrix} 0 & X_1^*Y_2 \\ -X_2^*Y_1 & 0 \end{bmatrix}\right\|_2 \\
&= \max\{\sigma_{\max}(X_1^*Y_2),\ \sigma_{\max}(X_2^*Y_1)\} \\
&= s_1.
\end{aligned}$$

但 $s_1^2 = 1 - c_1^2 = 1 - \sigma^2{}_{\min}(X_1^*Y_1)$，因此定理得证。

对于任意给定的两个子空间$\mathscr{X}$和$\mathscr{Y}$($\dim\mathscr{X} = \dim\mathscr{Y} = l$)，我们来看 $\mathrm{dist}(\mathscr{X}, \mathscr{Y})$的几何意义。从定理4.3的证明易知，存在$\mathscr{X}$和$\mathscr{Y}$的标准正交基$\{x_1, \cdots, x_l\}$和$\{y_1, \cdots, y_l\}$满足：

$$0 \leqslant x_1^*y_1 \leqslant x_2^*y_2 \leqslant \cdots \leqslant x_l^*y_l, \tag{4.7}$$

$$x_i^*y_j = 0, \quad i \neq j. \tag{4.8}$$

记 θ_i 为直线 $\mathrm{span}\{x_i\}$与 $\mathrm{span}\{y_i\}$之间夹角，即 $\theta_i = \arccos x_i^*y_i$，则有

$$\pi/2 \geqslant \theta_1 \geqslant \theta_2 \geqslant \cdots \geqslant \theta_l.$$

而且从定理 4.3 的证明亦可看出，这样得到的 l 个角度 $\theta_1, \theta_2, \cdots, \theta_l$ 与$\mathscr{X}$和$\mathscr{Y}$之满足(4.7)和(4.8)的基的选择无关，由$\mathscr{X}$和$\mathscr{Y}$唯一确定。因而我们可以称这些角的最大者 θ_1 为子空间 $\mathscr{X}$与$\mathscr{Y}$之间的夹角。这样，定理4.3就是说$\mathscr{X}$和$\mathscr{Y}$之间的距离$\mathrm{dist}(\mathscr{X}, \mathscr{Y})$正好是它们之间夹角的正弦，即

$$\mathrm{dist}(\mathscr{X}, \mathscr{Y}) = \sin\theta_1.$$

§5 非 负 矩 阵

5.1 基本概念和性质

定义5.1 设 $A = [a_{ij}] \in \mathbb{R}^{m\times n}$。如果 $a_{ij} \geqslant 0$ (或 $a_{ij} > 0$)

对所有的i,j成立，则称A是**非负（或正）矩阵**，记作$A\geqslant 0$（或$A>0$）。

设$A,B\in\mathbb{R}^{m\times n}$。如果$A-B\geqslant 0$（或$A-B>0$），则记作$A\geqslant B$（或$A>B$）。对于任意给定的$A=[a_{ij}]\in\mathbb{C}^{m\times n}$，我们用$|A|$来表示$A$的元素取绝对值之后所得到的非负矩阵，即$|A|=[|a_{ij}|]$；特别，当$x=(\alpha_1,\cdots,\alpha_n)^{\mathrm{T}}\in\mathbb{C}^n$时，$|x|=(|\alpha_1|,\cdots,|\alpha_n|)^{\mathrm{T}}$。

定理5.1（谱半径的单调性） 设$A\in\mathbb{C}^{n\times n}$，$B\in\mathbb{R}^{n\times n}$。如果$|A|\leqslant B$，则$\rho(A)\leqslant\rho(B)$。

证明 若不然，则有$\rho(A)>\rho(B)$。令$r=(\rho(A)+\rho(B))/2$，则$\rho(A)>r>\rho(B)$；再令$\tilde{A}=A/r$，$\tilde{B}=B/r$，则$\rho(\tilde{A})=\rho(A)/r>1$，$\rho(\tilde{B})=\rho(B)/r<1$。于是由定理3.8知，$\lim\limits_{k\to\infty}\tilde{B}^k=0$。而$|A|\leqslant B$蕴含着$|\tilde{A}^k|\leqslant|\tilde{A}|^k\leqslant\tilde{B}^k$对一切的自然数$k$成立，从而有$\lim\limits_{k\to\infty}\tilde{A}^k=0$。再应用定理3.8有$\rho(\tilde{A})<1$，这与$\rho(\tilde{A})>1$矛盾。因此，有$\rho(A)\leqslant\rho(B)$成立。

定义5.2 设$A\in\mathbb{R}^{n\times n}$。若存在$n$阶排列方阵$P$，使得

$$PAP^{\mathrm{T}}=\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \tag{5.1}$$

其中A_{11}和A_{22}是两个低阶方阵，则称A是**可分的**（或**可约的**）；否则称作**不可分的**（或**不可约的**）。

可分的概念来源于线性方程组的求解问题。一个线性方程组的系数矩阵是可分的，表明这一方程组，可通过适当调整方程和未知数的次序，化为两个低阶的方程组来求解。

定理5.2 设$A\in\mathbb{R}^{n\times n}$是非负的。则$A$不可分的充分必要条件是

$$(I+A)^{n-1}>0. \tag{5.2}$$

证明 *必要性* 假设A是不可分的。欲证(5.2)成立，只需证

$$(I+A)^{n-1}e_i>0, \quad i=1,2,\cdots,n \tag{5.3}$$

成立即可。

现任取 $x\in\mathbb{R}^n$，$x\geqslant 0$，$x\neq 0$。令 $x_0=x$，然后递推地定义

$$x_{k+1}=(I+A)x_k, \quad k=0,1,\cdots,n-2.$$

用 m_k 记 x_k 中非零分量的个数，显然有 $m_0\geqslant 1$，且 $m_{k+1}\geqslant m_k$，$k=0,1,\cdots,n-2$。这样欲证(5.3)成立，只需证 $m_{n-1}=n$ 即可。

如若 $m_{n-1}<n$，则必存在某个 k，使得 $m_k=m_{k+1}<n$。于是必存在一个排列方阵 P，使得

$$Px_{k+1}=\begin{bmatrix} u \\ 0 \end{bmatrix}, \quad Px_k=\begin{bmatrix} v \\ 0 \end{bmatrix}, \tag{5.4}$$

其中 $u,v\in\mathbb{R}^{m_k}$，且 $u>0$，$v>0$。由 $x_{k+1}=x_k+Ax_k$ 可得

$$Px_{k+1}=Px_k+PAP^{\mathrm{T}}Px_k. \tag{5.5}$$

对 PAP^{T} 作分块如下：

$$PAP^{\mathrm{T}}=\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\begin{matrix} m_k \\ n-m_k \end{matrix}. \tag{5.6}$$
$$\qquad\qquad m_k \qquad n-m_k$$

将(5.4)和(5.6)代入(5.5)，并注意到$u>0$，$v>0$，即可推出$A_{21}=0$。这与A不可分的假定矛盾。所以$m_{n-1}=n$成立，而这里的x是任取的，故(5.3)成立。

充分性　若A是可分的，即存在排列方阵P，使得

$$PAP^{\mathrm{T}}=\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix},$$

其中 A_{11} 和 A_{22} 是低阶方阵，从而有

$$P(I+A)^{n-1}P^{\mathrm{T}}=\begin{bmatrix} A_{11}+I & A_{12} \\ 0 & A_{22}+I \end{bmatrix}^{n-1}$$

$$=\begin{bmatrix} (A_{11}+I)^{n-1} & * \\ 0 & (A_{22}+I)^{n-1} \end{bmatrix},$$

这就是说$(I+A)^{n-1}$必含有零元素。因此，若$(I+A)^{n-1}>0$，则A必是不可分的。

5.2 Perron-Frobenius 定理

定理5.3(Perron-Frobenius 定理) 设$A\in\mathbb{R}^{n\times n}$是一个非负不可分矩阵。则

(1) $\rho(A)>0$，而且是A的一个单特征值；

(2) 对应于$\rho(A)$的特征向量可取作正的，即存在$x\in\mathbb{R}^n$，且$x>0$，使$Ax=\rho(A)x$；

(3) 不存在属于其他特征值的非负特征向量。

先证几个引理。

引理5.1 设$A\in\mathbb{R}^{n\times n}$是一个非负矩阵，$z\in\mathbb{R}^n$是一个不为零的非负向量。并假定$\xi\in\mathbb{R}$满足$Az>\xi z$，则$\rho(A)>\xi$。

证明 不妨设$\xi\geqslant0$。取$\varepsilon>0$满足

$$Az\geqslant(\xi+\varepsilon)z。\qquad(5.7)$$

令$B=A/(\xi+\varepsilon)$，则从(5.7)易得

$$B^kz\geqslant B^{k-1}z\geqslant\cdots\geqslant z\qquad(5.8)$$

对一切的自然数k成立。由$z\geqslant0$且$z\neq0$，从(5.8)知，当$k\to\infty$时，B^k不趋向于零，从而，据定理3.8，必有$\rho(B)\geqslant1$，即

$$\rho(A)\geqslant\xi+\varepsilon>\xi。$$

引理5.2 设$v_j\in\mathbb{C}$，$\alpha_j\in\mathbb{R}$，$\alpha_j>0$，$j=1,2,\cdots,m$。则

$$\left|\sum_{j=1}^{m}\alpha_jv_j\right|\leqslant\sum_{j=1}^{m}\alpha_j|v_j|,\qquad(5.9)$$

且等号成立的充分必要条件是存在$\eta\in\mathbb{C}$满足$|\eta|=1$，使得

$$\eta v_j\geqslant0,\quad j=1,2,\cdots,m,\qquad(5.10)$$

即$v_1,\cdots,v_m$位于同一条从原点出发的射线上。

证明 只对等号成立的必要条件加以证明.

当 $m=2$ 时，记 $v_j=r_je^{i\theta_j}$，$r_j\geqslant 0$，$0\leqslant\theta_j<2\pi$，$j=1,2$. 此时，(5.9)等号成立，即

$$|a_1r_1e^{i\theta_1}+a_2r_2e^{i\theta_2}|=a_1r_1+a_2r_2,$$

由此可得 $\cos(\theta_1-\theta_2)=1$，从而有 $\theta_1=\theta_2$，即 $m=2$ 时命题成立.

现假定命题对 $m=k-1$ 成立，我们来考虑 $m=k$ 的情形. 假定有

$$\left|\sum_{j=1}^{k}a_jv_j\right|=\sum_{j=1}^{k}a_j|v_j|$$

成立. 令 $v=\sum_{j=1}^{k-1}a_jv_j$，则上式蕴含着:

$$|v+a_kv_k|=|v|+a_k|v_k|, \tag{5.11}$$

$$\left|\sum_{j=1}^{k-1}a_jv_j\right|=\sum_{j=1}^{k-1}a_j|v_j|. \tag{5.12}$$

由 $m=2$ 所证的结果和归纳法假定知，存在 $0\leqslant\theta<2\pi$ 和 $0\leqslant\varphi<2\pi$，使

$$e^{i\theta}v\geqslant 0,\qquad e^{i\theta}v_k\geqslant 0, \tag{5.13}$$

$$e^{i\varphi}v_j\geqslant 0,\qquad j=1,2,\cdots,k-1. \tag{5.14}$$

从(5.14)可得

$$e^{i\varphi}v=\sum_{j=1}^{k-1}e^{i\varphi}v_j\cdot a_j\geqslant 0, \tag{5.15}$$

而 $0\leqslant\theta,\varphi<2\pi$，故(5.13)和(5.15)蕴含着 $\varphi=\theta$，因此命题对 $m=k$ 亦成立. 由归纳法原理知，命题对一切的自然数 m 成立.

引理5.3 设 $A=[a_{ij}]\in\mathbb{R}^{n\times n}$，且 $A>0$. 若 $u=(u_1,\cdots,u_n)^T\in\mathbb{C}^n$，$u\neq 0$，满足

$$Au=\lambda u,\qquad |\lambda|=\rho(A), \tag{5.16}$$

则有 $\lambda=\rho(A)>0$，$|u|>0$ 和 $A|u|=\rho(A)|u|$ 成立.

证明 先证对满足(5.16)的u，必存在单位复数η(即$|\eta|=1$)，使得

$$\eta u_j \geqslant 0, \quad j=1,2,\cdots,n. \tag{5.17}$$

若(5.17)不真，则由引理5.2知，必有

$$|\lambda u_k| = \left|\sum_{j=1}^{n} a_{kj} u_j\right| < \sum_{j=1}^{n} a_{kj} |u_j|$$

对一切的$k\in\{1,2,\cdots,n\}$成立，即有

$$|\lambda||u| < A|u|. \tag{5.18}$$

于是，根据引理5.1，应有$\rho(A)>|\lambda|$，这与$\rho(A)=|\lambda|$的假定矛盾，从而(5.17)成立。

从(5.17)成立，可知

$$|u| = \eta u.$$

因此，$|u|$也是属于λ的一个特征向量，即

$$A|u| = \lambda|u|. \tag{5.19}$$

由于$A>0$而$u\neq 0$，故(5.19)蕴含着$\lambda>0$和$|u|>0$，于是引理得证。

推论5.1 设$A\in\mathbb{R}^{n\times n}$，且$A>0$。则

(1) $\rho(A)$是A的正特征值；

(2) $\rho(A)$的几何重数是1，且对应的特征向量可取作正向量；

(3) 对任意的$\lambda\in\lambda(A)$，且$\lambda\neq\rho(A)$，必有$|\lambda|<\rho(A)$。

证明 除$\rho(A)$的几何重数为1之外，推论的其他结论都已包含在引理5.3之中。因此，下面只证$\rho(A)$的几何重数为1。

如若不然，则在$\mathbb{C}^n$中必存在两个线性无关的向量u和v，使得

$$Au=\rho(A)u, \qquad Av=\rho(A)v.$$

由于$u\neq 0$，故至少存在一个非零分量，不妨设它的第i个分量$u_i\neq 0$。令

$$z=v-\frac{v_i}{u_i}u,$$

其中v_i表示v的第i个分量，则z是一个至少有一个分量为零的非零向量，且亦是属于$\rho(A)$的特征向量。于是，由引理5.3知，应有$|z|>0$，而这又与z有一个分量为零矛盾，从而$\rho(A)$的几何重数是1。

引理5.4 设$A\in\mathbb{C}^{n\times n}$，$\lambda\in\lambda(A)$。则$\lambda$是$A$的单特征值的充分必要条件是：

(1) $\mathrm{rank}(A-\lambda I)=n-1$，即$\lambda$的几何重数是1；

(2) 属于λ的左右特征向量v和u满足$v^{\mathrm{T}}u\neq 0$。

证明 由于特征值的几何重数、代数重数以及条件(2)中所述的左右特征向量所满足的条件都在相似变换下保持不变，因此，不失一般性，可假定A是Jordan标准形，且对应于λ的Jordan块排在首位。此时，定理的必要性是显然的。因此，下面只证充分性。

由λ的几何重数为1，因此A具有如下形状

$$A=\begin{bmatrix}J_1 & 0\\ 0 & J_2\end{bmatrix},$$

其中J_2不含属于λ的Jordan块，而

$$J_1=\begin{bmatrix}\lambda & 1 & & 0\\ & \lambda & \ddots & \\ & & \ddots & 1\\ 0 & & & \lambda\end{bmatrix}$$

是属于λ的$m\times m$的Jordan块。这样，欲证λ的代数重数是1，只需证$m=1$即可。如若不然，则e_m和e_1分别是A属于λ的左右特征向量，而且$m>1$，从而$e_m^{\mathrm{T}}e_1=0$，这与条件(2)矛盾。因此$m=1$，

即λ的代数重数是1。

定理5.3的证明 分三步来证明。

(1) 先证：若$x\in\mathbb{C}^n$，$x\neq 0$，满足

$$Ax=\lambda x,\qquad |\lambda|=\rho(A),\tag{5.20}$$

则必有$|x|>0$，而且

$$A|x|=\rho(A)|x|.\tag{5.21}$$

如果(5.21)得证，则定理5.3的(2)得证，且有$\rho(A)>0$。

从(5.20)可得

$$\rho(A)|x|=|\lambda||x|\leqslant A|x|.\tag{5.22}$$

从(5.22)出发，归纳地可证

$$\rho(A)^k|x|\leqslant A^k|x|\tag{5.23}$$

对一切的自然数k成立。于是有

$$(1+\rho(A))^{n-1}|x|\leqslant(I+A)^{n-1}|x|.\tag{5.24}$$

由A是非负不可分的，故从定理5.2知，$(A+I)^{n-1}>0$，进而$(I+A^T)^{n-1}>0$。应用推论5.1于$(I+A^T)^{n-1}$上，可知存在$y>0$，使得

$$y^T(I+A)^{n-1}=\rho((I+A)^{n-1})y^T.\tag{5.25}$$

在(5.24)两边左乘y^T，并应用(5.25)，得

$$(1+\rho(A))^{n-1}y^T|x|\leqslant\rho((I+A)^{n-1})y^T|x|,\tag{5.26}$$

而$y^T|x|>0$，故有

$$(1+\rho(A))^{n-1}\leqslant\rho((I+A)^{n-1}).\tag{5.27}$$

另一方面由谱映照定理知，必存在$\mu\in\lambda(A)$，使得

$$\rho((I+A)^{n-1})=(1+\mu)^{n-1}.\tag{5.28}$$

将(5.28)代入(5.27)，并注意到幂函数的单调性，可得

$$1+\rho(A)\leqslant|1+\mu|\leqslant 1+|\mu|\leqslant 1+\rho(A)。$$

这表明，$\mu=|\mu|=\rho(A)$，从而(5.27)，进而(5.24)的等号必须成立，即

$$(1+\rho(A))^{n-1}|x|=(I+A)^{n-1}|x|。\qquad(5.29)$$

由$(I+A)^{n-1}>0$，且$|x|\neq 0$，知(5.29)蕴含着$|x|>0$。此外，从(5.29)和(5.23)可知，必有

$$A|x|=\rho(A)|x|,$$

即(5.21)成立。

(2) 再证：$\rho(A)$是A的单特征值。

由(1)所证知，对属于$\rho(A)$的任意特征向量u，必有$|u|>0$。因此，完全类似于推论5.1的证明，可证$\rho(A)$的几何重数是1。

另外，对A和A^{T}应用(1)所证，知属于$\rho(A)$的左右特征向量v和u可取作正的，因而有$v^{T}u>0$；于是，据引理5.4，知$\rho(A)$是A的单特征值。

(3) 最后证：不存在属于其他特征值的非负特征向量。

反证。如若不然，则存在不为零的非负向量z满足

$$Az=\lambda z,\qquad \lambda\neq\rho(A)。\qquad(5.30)$$

另一方面，从(1)所证知，存在$u>0$，使得

$$u^{T}A=\rho(A)u^{T}。\qquad(5.31)$$

在(5.30)两边左乘u^{T}，并注意到(5.31)，可得

$$\rho(A)u^{T}z=\lambda u^{T}z,$$

但$u^{T}z>0$，故有$\rho(A)=\lambda$，这与$\lambda\neq\rho(A)$的假定矛盾，从而(3)得证。

5.3 非负矩阵的谱

由推论5.1知，正矩阵A的模为$\rho(A)$的特征值是唯一的，而

非负不可分矩阵则不然，现看一个简单的例子。

例5.1 设

$$A=\begin{bmatrix} 0 & 1 & & & \\ 0 & 0 & 1 & & 0 \\ \vdots & \vdots & \ddots & \ddots & \\ 0 & 0 & & \ddots & 1 \\ 1 & 0 & \cdots & \cdots & 0 \end{bmatrix}\in \mathbb{R}^{n\times n}.$$

则可证A是非负不可分的，它的n个特征值是

$$\lambda_j=e^{i\frac{2j\pi}{n}},\ j=0,1,\cdots,n-1,$$

从而它的特征值的模都等于谱半径$\rho(A)=1$。

定理5.4 设A是n阶非负不可分矩阵，h是A的模等于$\rho(A)$的不同特征值的个数．则

(1) A的模为$\rho(A)$的h个特征值是

$$\lambda_j=\rho(A)e^{i\frac{2i\pi}{h}},\ j=0,\ \ 1,\cdots,h-1,$$

也就是说，它们“均匀”地分布在以原点为圆心，$\rho(A)$为半径的圆周上；

(2) A的特征多项式具有如下形状：

$p(t)=t^m[t^h-\rho(A)^h][t^h-\delta_2\rho(A)^h]\cdots[t^h-\delta_r\rho(A)^h]$,

其中$rh+m=n$，且当$r>1$时，$0<|\delta_i|<1$，$i=2,\cdots,r$；即除模为$\rho(A)$的特征值外，A的其余非零特征值亦可分为若干组，使得每组正好有A的h个模相等的特征值，且它们“均匀”地分布在某一圆心在原点半径小于$\rho(A)$的圆周上。

这一定理的证明较繁，由于篇幅所限，这里不再给出，有兴趣的读者可参看文献[69]。

例5.2 设

$$A=\begin{bmatrix}0&0&1&0\\0&0&1&1\\0&1&0&0\\1&1&0&0\end{bmatrix}.$$

则A是非负不可分的，它的特征值是

$$\lambda=\pm\sqrt{1+\sqrt{2}},\qquad \lambda=\pm i\sqrt{\sqrt{2}-1}.$$

其特征值的分布情况如图 5.1 所示。

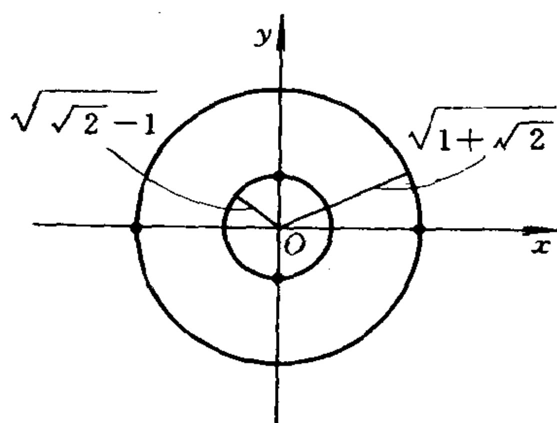


图5.1　黑点表示特征值所在位置

从可分的定义5.2，容易证明

定理5.5　设A是一个n阶非负矩阵，则存在一个排列方阵P，使得

$$PAP^{\mathrm{T}}=\begin{bmatrix}A_{11}&A_{12}&\cdots&A_{1k}\\&A_{22}&\cdots&A_{2k}\\&0&\ddots&\vdots\\&&&A_{kk}\end{bmatrix},\tag{5.32}$$

其中$A_{ii}(i=1,2,\cdots,k)$是非负不可分方阵。

结合定理5.3—5.5，可得

推论5.2　设A是一个n阶非负矩阵。则

(1) $\rho(A)$是A的特征值，且属于$\rho(A)$的特征向量可取作非

负的，即存在不为零的非负向量x，使得$Ax=\rho(A)x$；

(2) A的特征值可分成若干组，每组中的特征值模都相等，而且“均匀”地分布在以原点为心的某一圆周上；

(3) 若A有一个正的特征向量x，则x必是属于$\rho(A)$的特征向量。

5.4 Birkhoff 定理

定义5.3 设$A=[a_{ij}]\in\mathbb{R}^{n\times n}$。如果$A$满足

$$a_{ij}\geqslant 0,\ \sum_{i=1}^{n}a_{ij}=\sum_{j=1}^{n}a_{ij}=1,\ i,j=1,2,\cdots,n,$$

则称A是**双随机阵**。

从上述定义易知，若A是双随机阵，则$\rho(A)=1$，而且对应的正特征向量可取作$e=(1,1,\cdots,1)^{\mathrm{T}}$。此外，任一排列方阵必是双随机阵；反之，一个双随机阵是排列方阵的充分与必要条件是它只有n个非零元素。

定理5.6(Birkhoff 定理) 所有n阶双随机阵的集合是所有n阶排列方阵的凸包；即任一n阶双随机阵$A=[a_{ij}]$，必可表示成n阶排列方阵$P_i(i=1,2,\cdots,n!)$的凸组合：

$$A=\sum_{i=1}^{n!}\sigma_i P_i,\quad \sum_{i=1}^{n!}\sigma_i=1,\ \sigma_i\geqslant 0,\quad i=1,\cdots,n!.$$

证明 对A的非零元素的个数$\nu(A)$用数学归纳法。显然有$\nu(A)\geqslant n$。

当$\nu(A)=n$时，A就是排列方阵，因而定理对$\nu(A)=n$的情形自然成立。

现假定$\nu(A)>n$，而且假定对于满足$\nu(B)<\nu(A)$的双随机阵B已证定理成立。

由于$\nu(A)>n$，故A至少有一行含有两个以上的非零元素，从而必存在指标i_1和j_1，使得$0<a_{i_1j_1}<1$；于是在第j_1列必然有

另一个非零元素，即存在$i_2 \neq i_1$，使得$0<a_{i_2 j_1}<1$；同样在第i_2行又有另一个非零元$a_{i_2 j_2}$满足$0<a_{i_2 j_2}<1$，$j_2 \neq j_1$。如此下去，我们就可从$a_{i_1 j_1}$出发，找到一列$a_{i_k j_k}$和$a_{i_k j_{k-1}}$满足

$$0<a_{i_k j_k},\ a_{i_k j_{k-1}}<1, \quad k=2,3,\cdots.$$

而A的行数是一个有限数n，故到某一步所得到的行指标i_{s+1}必与前面某一步得到的行指标i_t相重，即$i_{s+1}=i_t$。现不妨假定$t=1$，否则可将出发点移到第i_t行，然后重新编号即可。这表明，我们可以找到s个互不相同的行指标$i_1,i_2,\cdots,i_s$和s个列指标$j_1,\cdots,j_s$满足

$$0<a_{i_k j_k},\ a_{i_{k+1}, j_k}<1,\ k=1,2,\cdots,s,$$

其中$i_{s+1}=i_1$。

将这$2s$个元素所在的位置分为两组：

$$\mathscr{L}_1=\{(i_k,j_k);\ k=1,2,\cdots,s\},$$
$$\mathscr{L}_2=\{(i_1,j_s),(i_k,j_{k-1});\ k=2,\cdots,s\}$$

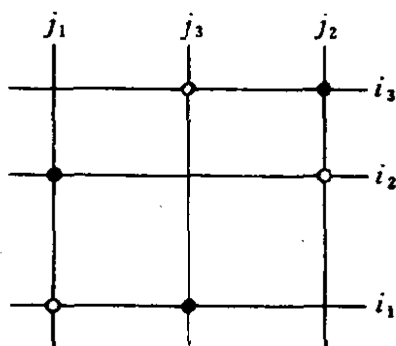


图5.2 ○属于$\mathscr{L}_1$，●属于$\mathscr{L}_2$

图5.2给出了$s=3$时，所找到的6个元素所在位置示意图，并标出了其分组情况。

令

$$\alpha=\min_{(i,j)\in\mathscr{L}_1} a_{ij}, \quad \beta=\min_{(i,j)\in\mathscr{L}_2} a_{ij},$$

$$\hat{a}_{ij}=\begin{cases} a_{ij}-\alpha, & (i,j)\in\mathscr{L}_1, \\ a_{ij}+\alpha, & (i,j)\in\mathscr{L}_2, \\ a_{ij}, & \text{其他}, \end{cases}$$

$$
\tilde{a}_{ij}=\begin{cases}a_{ij}+\beta, & (i,j)\in\mathscr{L}_1,\\ a_{ij}-\beta, & (i,j)\in\mathscr{L}_2,\\ a_{ij}, & \text{其他}.\end{cases}
$$

记 $\hat{A}=[\hat{a}_{ij}]$，$\tilde{A}=[\tilde{a}_{ij}]$，则 $\hat{A}$ 和 $\tilde{A}$ 都是双随机阵，且 $\nu(\hat{A})<\nu(A)$，$\nu(\tilde{A})<\nu(A)$。于是由归纳法假设，有

$$
\hat{A}=\sum_{i=1}^{n!}\hat{\sigma}_i P_i,\qquad \hat{\sigma}_i\geqslant 0,\quad \sum_{i=1}^{n!}\hat{\sigma}_i=1,
$$

$$
\tilde{A}=\sum_{i=1}^{n!}\tilde{\sigma}_i P_i,\qquad \tilde{\sigma}_i\geqslant 0,\quad \sum_{i=1}^{n!}\tilde{\sigma}_i=1.
$$

注意到

$$
A=\frac{\beta}{\alpha+\beta}\hat{A}+\frac{\alpha}{\alpha+\beta}\tilde{A},
$$

令

$$
\sigma_i=\frac{\beta}{\alpha+\beta}\hat{\sigma}_i+\frac{\alpha}{\alpha+\beta}\tilde{\sigma}_i,
$$

就有

$$
A=\sum_{i=1}^{n!}\sigma_i P_i,\qquad \sigma_i\geqslant 0,\quad \sum_{i=1}^{n!}\sigma_i=1.
$$

由归纳法原理知定理得证。

§6 有关矩阵特征值的几个重要定理

6.1 一般方阵的 Bauer-Fike 定理

首先，我们叙述一个非常有用的关于特征值的包含定理——Gerschgorin圆盘定理。它的证明在许多数值代数的教课书中都可找到，因此这里不再赘述，有兴趣的读者可查阅有关的文献，例如可参看[10]。

定理6.1(Gerschgorin 定理) 设 $A=[a_{ij}]\in\mathbb{C}^{n\times n}$。令

$$G_i(A)=\left\{z\in\mathbb{C}:\ |z-a_{ii}|\leqslant\sum_{j\neq i}|a_{ij}|\right\},\quad i=1,2,\cdots,n. \tag{6.1}$$

则

(1) $$\lambda(A)\subset\bigcup_{i=1}^{n}G_i(A); \tag{6.2}$$

(2) 如果(6.1)所定义的圆盘中，有m个互相连通且与其余$n-m$个不连通，则在此m个圆盘所成的连通区域中，恰有A的m个特征值。

例6.1 设

$$A=\begin{bmatrix}1 & 1 & 0\\ 1/8 & 2 & 1/8\\ 0 & 1/3 & 5\end{bmatrix}.$$

A的三个 Gerschgorin 圆盘分别为：

$$G_1(A)=\{z\in\mathbb{C}:\ |z-1|\leqslant 1\},$$
$$G_2(A)=\{z\in\mathbb{C}:\ |z-2|\leqslant 1/4\},$$
$$G_3(A)=\{z\in\mathbb{C}:\ |z-5|\leqslant 1/3\}.$$

容易看出，$G_1(A)$与$G_2(A)$相交，而$G_3(A)$与另外两个圆盘分离。因此，由 Gerschgorin 定理，我们可以断言，在$G_3(A)$内必有A的一个特征值，而在$G_1(A)\cup G_2(A)$内有A的两个特征值。如果我们再取对角阵$D=\mathrm{diag}(1,1,1/16)$，则由

$$DAD^{-1}=\begin{bmatrix}1 & 1 & 0\\ 1/8 & 2 & 2\\ 0 & 1/48 & 5\end{bmatrix}$$

可判定出A有一个特征值λ满足

$$|\lambda-5|<1/48.$$

例6.1表明，灵活应用 Gerschgorin 定理，有时会给出特征值的非常好的估计。

下面我们给出一个一般方阵的特征值的扰动定理。

定理6.2(Bauer-Fike定理) 设$A,B\in\mathbb{C}^{n\times n}$，其中$A$可对角化，即$A=Q^{-1}\Lambda Q$，$\Lambda$为对角矩阵，$Q$为非奇异矩阵。则对任意的$\mu\in\lambda(B)$，必存在$\lambda\in\lambda(A)$，使得

$$|\lambda-\mu|\leqslant\|Q^{-1}\|_2\|Q\|_2\|A-B\|_2. \tag{6.3}$$

证明 若$\mu\in\lambda(A)$，则(6.3)自然成立。下面考虑$\mu\notin\lambda(A)$的情形。设$x\in\mathbb{C}^n$是B之属于μ的特征向量，即$Bx=\mu x$，于是有

$$(B-A)x=\mu x-Ax=(\mu I-A)x.$$

但由于$\mu\notin\lambda(A)$，故$(\mu I-A)$可逆，从而有

$$x=(\mu I-A)^{-1}(B-A)x=Q^{-1}(\mu I-\Lambda)^{-1}Q(B-A)x.$$

上式两边取2范数，并利用矩阵谱范数的相容性和$\|x\|_2>0$，有

$$1\leqslant\|Q\|_2\|Q^{-1}\|_2\|A-B\|_2(\min_{\lambda\in\lambda(A)}|\lambda-\mu|)^{-1}$$

即

$$\min_{\lambda\in\lambda(A)}|\lambda-\mu|\leqslant\|Q\|_2\|Q^{-1}\|_2\|A-B\|_2,$$

从而有(6.3)成立。证毕。

1982年，Kahan,Parlett和蒋尔雄，将上述定理推广到一般方阵，他们的证明需要下面的结果。

引理6.1 设对角元素为λ的k阶若当块J的奇异值为$\sigma_1\geqslant\sigma_2\geqslant\cdots\geqslant\sigma_k>0$。则

$$\sigma_k\geqslant\frac{|\lambda|^k}{(1+|\lambda|)^{k-1}}. \tag{6.4}$$

证明 由假定知矩阵

$$T=J^*J=\begin{bmatrix}\bar\lambda & & & \\ 1 & \bar\lambda & & 0 \\ & \ddots & \ddots & \\ 0 & & 1 & \bar\lambda\end{bmatrix}\begin{bmatrix}\lambda & 1 & & 0 \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda\end{bmatrix}$$

$$
=\begin{bmatrix} |\lambda|^2 & \bar{\lambda} & & 0 \\ \lambda & 1+|\lambda|^2 & \ddots & \\ & \ddots & \ddots & \bar{\lambda} \\ 0 & & \lambda & 1+|\lambda|^2 \end{bmatrix}
$$

的特征值为 $\sigma_1^2 \geqslant \sigma_2^2 \geqslant \cdots \geqslant \sigma_k^2 > 0$。利 用 Gerschgorin 定理，可得

$$
0 < \sigma_i^2 \leqslant 1 + |\lambda|^2 + 2|\lambda|, \qquad i = 1, 2, \cdots, k.
$$

于是

$$
\sigma_k^2 = \frac{\prod_{i=1}^{k} \sigma_i^2}{\prod_{i=1}^{k-1} \sigma_i^2} = \frac{\det(T)}{\prod_{i=1}^{k-1} \sigma_i^2} \geqslant \frac{|\lambda|^{2k}}{(1+|\lambda|)^{2(k-1)}},
$$

由此立得(6.4)。

定理6.3(广义Bauer-Fike定理) 设 $A, B \in \mathbb{C}^{n\times n}$，并假定 A 的Jordan分解为 $A = Q^{-1}JQ$，其中 J 是 A 的Jordan标准形。则对任意 $\mu \in \lambda(B)$，必有 $\lambda \in \lambda(A)$ 使得

$$
\frac{|\lambda - \mu|^m}{(1 + |\lambda - \mu|)^{m-1}} \leqslant \|Q(A-B)Q^{-1}\|_2, \qquad (6.5)
$$

其中 m 是 J 中属于 λ 的最大 Jordan 块的阶数。

证明 无妨假设 $\mu \notin \lambda(A)$。此时，$\mu I - J$ 可逆。完全类 似于 Bauer-Fike 定理的证明可证

$$
1 \leqslant \|(\mu I - J)^{-1}\|_2 \|Q(A-B)Q^{-1}\|_2. \qquad (6.6)
$$

设 $J = \operatorname{diag}(J_1, J_2, \cdots, J_r)$，其中 J_i 是以 A 的特征值 λ_i 为对角元素的 k_i 阶 Jordan 块，$i = 1, 2, \cdots, r$。则由引理6.1可得

$$
\|(\mu I - J)^{-1}\|_2^{-1} = \sigma_{\min}(\mu I - J) = \min_{1<i<r} \sigma_{\min}(\mu I - J_i)
$$

$$
\geqslant \min_{1<i<r} \frac{|\mu - \lambda_i|^{k_i}}{(1 + |\mu - \lambda_i|)^{k_i - 1}}, \qquad (6.7)
$$

其中 $\sigma_{\min}(\mu I-J)$ 表示 $(\mu I-J)$的最小奇异值。注意到函数 $a^k/(1+a)^{k-1}$ 对任意给定的 $a>0$，随着 k 的增加而减小，就知必存在 A 的某个特征值 λ，使得

$$\min_{1\leqslant i\leqslant r}\frac{|\mu-\lambda_i|^{k_i}}{(1+|\mu-\lambda_i|)^{k_i-1}}=\frac{|\mu-\lambda|^m}{(1+|\mu-\lambda|)^{m-1}}, \quad (6.8)$$

其中 m 是 J 中属于 λ 的 Jordan 块的最大阶数。由 (6.8)，(6.7) 和 (6.6) 立即知定理的结论成立。

6.2 正规矩阵的 Hoffman-Wielandt 定理

利用正规矩阵可酉对角化这一特征，可以得到一些非常漂亮的结果。这里，先证一个简单而又十分有用的结果。

定理6.4 设 $A\in\mathbb{C}^{n\times n}$ 是正规矩阵。任取 $x\in\mathbb{C}^n$, $\|x\|_2=1$，定义 $\mu(x)=x^*Ax$，则

$$\min_{\lambda\in\lambda(A)}|\lambda-\mu(x)|\leqslant\|(A-\mu(x)I)x\|_2. \quad (6.9)$$

证明 由于 A 是正规矩阵，故存在 $U\in\mathscr{U}_n$，使得 $A=U\Lambda U^*$，其中 $\Lambda=\mathrm{diag}(\lambda_1,\cdots,\lambda_n)$。于是

$$\begin{aligned}\|(A-\mu(x)I)x\|_2&=\|U(\Lambda-\mu(x)I)U^*x\|_2\\&=\|(\Lambda-\mu(x)I)U^*x\|_2\\&\geqslant\min_{1\leqslant i\leqslant n}|\lambda_i-\mu(x)|\,\|U^*x\|_2\\&=\min_{1\leqslant i\leqslant n}|\lambda_i-\mu(x)|,\end{aligned}$$

即定理得证。

在定理 6.4 中取 $x=e_i(i=1,2,\cdots,n)$即得

推论6.1 设 $A=[a_{ij}]\in\mathbb{C}^{n\times n}$是正规矩阵。令

$$\mathscr{D}_i(A)=\left\{z\in\mathbb{C}:\ |z-a_{ii}|\leqslant\left(\sum_{j\neq i}|a_{ij}|^2\right)^{1/2}\right\}, i=1,\cdots,n.$$

则每个圆盘 $\mathscr{D}_i(A)$内至少有 A 的一个特征值。

注6.1 定理 6.4 通常称作 Krylov-Bogoljubov-Weinstein 定

理。尽管它的证明特别简单，但它揭示了正规矩阵的一个重要性质：任一单位向量x，数$\mu(x)=x^*Ax$可作为A的某个特征值的近似估计，其精确程度可以用剩余向量$r(x)=Ax-\mu(x)x$的2范数的大小来恒量。

定理6.5（Hoffman-Wielandt 定理） 设A和B是两个n阶正规矩阵，它们的特征值分别是$\lambda_1,\lambda_2,\cdots,\lambda_n$和$\mu_1,\mu_2,\cdots,\mu_n$。则存在$1,2,\cdots,n$的一个排列$\pi(1),\pi(2),\cdots,\pi(n)$，使得

$$\left(\sum_{i=1}^{n}|\mu_{\pi(i)}-\lambda_i|^2\right)^{1/2}\leqslant\|B-A\|_F. \tag{6.10}$$

证明 设A和B的 Schur 分解为

$$A=U\Lambda U^*,\quad B=V\Omega V^*, \tag{6.11}$$

其中$U,V\in\mathscr{U}_n$，$\Lambda=\mathrm{diag}(\lambda_1,\cdots,\lambda_n)$，$\Omega=\mathrm{diag}(\mu_1,\cdots,\mu_n)$。于是

$$\begin{aligned}\|B-A\|_F^2&=\mathrm{tr}([V\Omega V^*-U\Lambda U^*][V\Omega V^*-U\Lambda U^*]^*)\\&=\sum_{i=1}^{n}|\mu_i|^2+\sum_{i=1}^{n}|\lambda_i|^2\\&\quad-\mathrm{tr}(\Lambda U^*V\bar{\Omega}V^*U+U^*V\Omega V^*U\bar{\Lambda}).\end{aligned} \tag{6.12}$$

令

$$W=U^*V=[\omega_{ij}],\quad \theta_{ij}=\lambda_i\bar{\mu}_j+\bar{\lambda}_i\mu_j,\quad 1\leqslant i,j\leqslant n.$$

直接计算可知

$$\mathrm{tr}(\Lambda U^*V\bar{\Omega}V^*U+U^*V\Omega V^*U\bar{\Lambda})=\sum_{i,j=1}^{n}\theta_{ij}|\omega_{ij}|^2. \tag{6.13}$$

由于$W\in\mathscr{U}_n$，故$S_0=[|\omega_{ij}|^2]$是一双随机阵。下面我们的主要任务就是给出（6.13）的上界估计。为此，我们来考虑如下定义的线性泛函

$$f(S)=\sum_{i,j=1}^{n}\theta_{ij}\sigma_{ij}, \tag{6.14}$$

其中$S=[\sigma_{ij}]\in\mathbb{C}^{n\times n}$是双随机阵。

根据 Birkhoff 定理，对于任意的双随机阵S，有

$$S=\sum_{i=1}^{n!}\sigma_i P_i,\quad \sigma_i\geqslant 0,\quad \sum_{i=1}^{n!}\sigma_i=1,$$

其中 $P_1,\cdots,P_{n!}$ 是所有不同的 n 阶排列方阵。令排列方阵 P 满足

$$f(P)=\max_{1\leqslant i\leqslant n!} f(P_i),$$

则

$$f(S)=\sum_{i=1}^{n!}\sigma_i f(P_i)\leqslant f(P) \tag{6.15}$$

对一切的双随机阵 S 成立。

而对于排列方阵 P，必存在 $1,2,\cdots,n$ 的一个排列 $\pi(1)$，$\pi(2),\cdots,\pi(n)$，使得

$$e_i^{\mathrm{T}}P=e_{\pi(i)}^{\mathrm{T}},\quad i=1,2,\cdots,n,$$

于是

$$f(P)=\sum_{i=1}^{n}\theta_{i\pi(i)}. \tag{6.16}$$

将（6.16）代入（6.15),得

$$f(S)\leqslant\sum_{i=1}^{n}\theta_{i\pi(i)}$$

对一切的双随机阵 S 成立；特别，对 $S_0=[|\omega_{ij}|^2]$有

$$f(S_0)=\sum_{i,j=1}^{n}\theta_{ij}|\omega_{ij}|^2\leqslant\sum_{i=1}^{n}\theta_{i\pi(i)}$$

$$=\sum_{i=1}^{n}(\lambda_i\bar{\mu}_{\pi(i)}+\bar{\lambda}_i\mu_{\pi(i)}). \tag{6.17}$$

从(6.12),(6.13)和(6.17)得

$$\|B-A\|_F^2\geqslant\sum_{i=1}^{n}(|\mu_i|^2+|\lambda_i|^2-\lambda_i\bar{\mu}_{\pi(i)}-\bar{\lambda}_i\mu_{\pi(i)})$$

$$=\sum_{i=1}^{n}|\lambda_i-\mu_{\pi(i)}|^2,$$

即定理得证。

推论6.2 设 $A,B\in\mathbb{C}^{n\times n}$ 是 Hermite 矩阵，它们的特征值分别为

$$\lambda_1\geqslant\cdots\geqslant\lambda_n \text{ 和 } \mu_1\geqslant\cdots\geqslant\mu_n.$$

则

$$\left[\sum_{i=1}^{n}(\lambda_i-\mu_i)^2\right]^{\frac{1}{2}}\leqslant\|B-A\|_F.$$

证明 只需证：对于 $1,2,\cdots,n$ 的任一排列 $\pi(1),\pi(2),\cdots,\pi(n)$，都有

$$\sum_{i=1}^{n}(\lambda_i-\mu_i)^2\leqslant\sum_{i=1}^{n}(\lambda_i-\mu_{\pi(i)})^2$$

即可。详细证明留作练习。

推论6.3 设 $A,B\in\mathbb{C}^{m\times n}$，它们的奇异值分别是

$$\sigma_1\geqslant\cdots\geqslant\sigma_n\geqslant0 \text{ 和 } \tau_1\geqslant\cdots\geqslant\tau_n\geqslant0.$$

则

$$\left(\sum_{i=1}^{n}(\sigma_i-\tau_i)^2\right)^{1/2}\leqslant\|B-A\|_F. \tag{6.18}$$

证明 不失一般性，可假定 $m=n$。设 A 的奇异值分解为 $A=U\Sigma V^*$，其中 $U,V\in\mathscr{U}_n,\Sigma=\mathrm{diag}(\sigma_1,\cdots,\sigma_n)$。$\sigma_1\geqslant\cdots\geqslant\sigma\geqslant0$。直接验证可知

$$\begin{bmatrix}0 & A\\ A^* & 0\end{bmatrix}=\hat{U}\begin{bmatrix}\Sigma & 0\\ 0 & -\Sigma\end{bmatrix}\hat{U}^*,$$

其中

$$\hat{U}=\sqrt{\frac{1}{2}}\begin{bmatrix}U & U\\ V & -V\end{bmatrix}$$

为 $2n$ 阶酉矩阵。从而$\begin{bmatrix}0 & A\\ A^* & 0\end{bmatrix}$的特征值为

$$\sigma_1\geqslant\cdots\geqslant\sigma_n\geqslant-\sigma_n\geqslant\cdots\geqslant-\sigma_1.$$

同理可证$\begin{bmatrix} 0 & B \\ B^* & 0 \end{bmatrix}$的特征值为

$$\tau_1 \geqslant \cdots \geqslant \tau_n \geqslant -\tau_n \geqslant \cdots \geqslant -\tau_1.$$

对 Hermite 矩阵$\begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix}$和$\begin{bmatrix} 0 & B \\ B^* & 0 \end{bmatrix}$应用推论6.2，可得

$$2\sum_{i=1}^{n} (\tau_i - \sigma_i)^2 \leqslant \left\| \begin{bmatrix} 0 & B \\ B^* & 0 \end{bmatrix} - \begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix} \right\|_F^2$$

$$= 2\|B - A\|_F^2.$$

由此即知（6.18）成立。

6.3 Hermite 矩阵的极小极大定理

定理6.6（Courant-Fischer极小极大定理） 设 $A \in \mathbb{C}^{n \times n}$ 为 Hermite 矩阵，其特征值为 $\lambda_1 \geqslant \cdots \geqslant \lambda_n$。则有

$$\lambda_i = \max_{\mathscr{X} \in \mathbb{G}_i^n} \min_{u \in S(\mathscr{X})} u^* A u \qquad (6.19)$$

$$= \min_{\mathscr{X} \in \mathbb{G}_{n-i+1}^n} \max_{u \in S(\mathscr{X})} u^* A u, \qquad (6.20)$$

其中 $\mathbb{G}_l^n$ 表示 $\mathbb{C}^n$ 中所有 l 维子空间的全体，

$$S(\mathscr{X}) = \{u \in \mathscr{X} \mid \|u\|_2 = 1\}, \quad i = 1, 2, \cdots, n.$$

证明 先证（6.19）成立。设 $u_1, \cdots, u_n$ 为对应于 $\lambda_1, \cdots, \lambda_n$ 的特征向量所构成的 $\mathbb{C}^n$ 的一组标准正交基。

现任取$\mathbb{C}^n$ 的一个 i 维子空间$\mathscr{X}$，并假定 $x_1, x_2, \cdots, x_i$ 是$\mathscr{X}$ 的一组基。将每个 x_j 按 $u_1, \cdots, u_n$ 展开有

$$x_j = \sum_{k=1}^{n} \alpha_{kj} u_k, \quad j = 1, 2, \cdots, i. \qquad (6.21)$$

记 $X = [x_1, \cdots, x_i]$，$U = [u_1, \cdots, u_n]$，$B = [\alpha_{kj}]$，则(6.21)写成矩阵形式即为 $X = UB$。现将 B 作如下分块

$$B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \begin{matrix} i-1 \\ n-i+1, \end{matrix}$$
$$\quad\; i$$

则齐次方程组

$$B_1 y = 0 \tag{6.22}$$

有 $i-1$ 个方程，i 个未知数，故必有非零解；又 X 的列是线性无关的，故对(6.22)的任意非零解 y 都有 $B_2 y \neq 0$。因此，可取(6.22)的一个非零解 y_0，使得 $\|B_2 y_0\|_2 = 1$。记 $B_2 y_0 = (\xi_i, \xi_{i+1}, \cdots, \xi_n)^{\mathrm{T}}$，则 $x_0 = X y_0 = U B y_0 = \sum\limits_{k=i}^{n} \xi_k u_k \in \mathscr{X}$，且 $\|x_0\|_2 = \|B y_0\|_2 = \|B_2 y_0\|_2 = 1$。因而有

$$\begin{aligned}\min_{u \in S(\mathscr{X})} u^* A u &\leqslant x_0^* A x_0 = \left(\sum_{k=i}^{n} \xi_k u_k\right)^* A \left(\sum_{k=i}^{n} \xi_k u_k\right) \\ &= \sum_{k=i}^{n} |\xi_k|^2 \lambda_k \leqslant \lambda_i \sum_{k=i}^{n} |\xi_k|^2 = \lambda_i.\end{aligned} \tag{6.23}$$

另一方面，对于 $\mathscr{X}_0 = \mathrm{span}\{u_1, \cdots, u_i\}$ 这一特殊的 i 维子空间，有

$$\min_{u \in S(\mathscr{X}_0)} u^* A u = \min_{\sum\limits_{k=1}^{i} |\xi_k|^2 = 1} \sum_{k=1}^{i} |\xi_k|^2 \lambda_k = \lambda_i. \tag{6.24}$$

由（6.23）和（6.24）即知（6.19）成立。

至于（6.20），只需应用（6.19）于 $-A$ 上即可。证毕。

下面我们利用 Hermite 矩阵的极小极大定理来证明几个十分重要的结果。首先分别在（6.19）和（6.20）中取 i 为1和 n，就得到

推论6.4 设 n 阶 Hermite 矩阵 A 的最大与最小特征值分别为 λ_1 与 λ_n。则

$$\lambda_1 = \max_{\substack{u \in \mathbf{C}^n \\ \|u\|_2 = 1}} u^* A u, \quad \lambda_n = \min_{\substack{u \in \mathbf{C}^n \\ \|u\|_2 = 1}} u^* A u.$$

定理6.7(分隔定理) 设A为n阶 Hermite 矩阵，$B=U^*AU$，其中$U\in\mathbb{C}^{n\times(n-1)}$满足$U^*U=I_{n-1}$。再设$A$与$B$的特征值分别为$\lambda_1\geqslant\cdots\geqslant\lambda_n$和$\mu_1\geqslant\cdots\geqslant\mu_{n-1}$。则

$$\lambda_1\geqslant\mu_1\geqslant\lambda_2\geqslant\mu_2\geqslant\cdots\geqslant\mu_{n-1}\geqslant\lambda_n. \tag{6.25}$$

证明 由极小极大定理知，存在一个i维子空间$\mathscr{Y}\subset\mathbb{C}^{n-1}$，使得

$$\mu_i=\min_{v\in S(\mathscr{Y})} v^*Bv. \tag{6.26}$$

令

$$\mathscr{X}=\{Uy\mid y\in\mathscr{Y}\},$$

则易证$\mathscr{X}$是$\mathbb{C}^n$中的一个i维子空间。于是

$$\begin{aligned}\mu_i&=\min_{v\in S(\mathscr{Y})} v^*Bv=\min_{v\in S(\mathscr{Y})} v^*U^*AUv\\&=\min_{u\in S(\mathscr{X})} u^*Au\leqslant\max_{\mathscr{X}\in\mathscr{G}_i^n}\min_{u\in S(\mathscr{X})} u^*Au\\&=\lambda_i.\end{aligned} \tag{6.27}$$

完全类似于(6.27)的证明，利用定理6.6的(6.20)，可证

$$\mu_i\geqslant\lambda_{i+1}. \tag{6.28}$$

由(6.27)和(6.28)即得(6.25)。定理证毕。

从定理6.7很容易导出下面一个常用的结果。

推论6.5 设$A\in\mathbb{C}^{n\times n}$为 Hermite 矩阵，$B$是$A$的一个$k$阶主子阵$(1\leqslant k\leqslant n-1)$。并设$A$与$B$的特征值分别为$\lambda_1\geqslant\cdots\geqslant\lambda_n$和$\mu_1\geqslant\cdots\geqslant\mu_k$。则

$$\lambda_i\geqslant\mu_i\geqslant\lambda_{n-k+i},\quad i=1,2,\cdots,k.$$

证明留作练习。

定理6.8 设n阶 Hermite 矩阵A与B的特征值分别为$\lambda_1\geqslant\cdots\geqslant\lambda_n$和$\mu_1\geqslant\cdots\geqslant\mu_n$。并设$E=B-A$的最大与最小特征值分别为$\varepsilon_1$和$\varepsilon_n$。则

$$\lambda_i + \varepsilon_n \leqslant \mu_i \leqslant \lambda_i + \varepsilon_1, \quad i = 1, 2, \cdots, n. \tag{6.29}$$

证明 首先对矩阵A应用定理 6.6 知，存在一个 $n-i+1$ 维子空间 $\mathscr{X} \subset \mathbb{C}^n$，使得

$$\lambda_i = \max_{u \in S(\mathscr{X})} u^* A u.$$

然后再对B和E应用定理 6.6 和推论 6.4 知，对上面所得到的子空间$\mathscr{X}$，有

$$\mu_i \leqslant \max_{u \in S(\mathscr{X})} u^* B u \leqslant \max_{u \in S(\mathscr{X})} u^* A u + \max_{u \in S(\mathscr{X})} u^* E u$$
$$= \lambda_i + \max_{u \in S(\mathscr{X})} u^* E u \leqslant \lambda_i + \varepsilon_1. \tag{6.30}$$

其次，在上述证明过程中交换A与B的位置，并注意到$-E$的最大特征值为$-\varepsilon_n$，即有

$$\lambda_i \leqslant \mu_i - \varepsilon_n. \tag{6.31}$$

由（6.30）和（6.31）可知（6.29）成立。定理证毕。

由定理 6.8 立即可得下述重要结果。

定理6.9（Weyl 定理） 在定理 6.8 的假设下，有

$$|\mu_i - \lambda_i| \leqslant \|A - B\|_2, \quad i = 1, 2, \cdots, n.$$

推论6.6 在推论 6.3 的假设下，有

$$|\tau_i - \sigma_i| \leqslant \|A - B\|_2, \quad i = 1, 2, \cdots, n.$$

习　题

1. 设

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

其中A_{ij}均为方阵，并满足$A_{11}A_{21} = A_{21}A_{11}$。证明

$$\det(A) = \det(A_{11}A_{22} - A_{21}A_{12}).$$

2. 设H_1为正定矩阵，H_2是与H_1同阶的 Hermite 矩阵。试证：$H_1 + H_2$为正定矩阵的充分必要条件是$H^{-1}H_2$的特征值均

大于 -1.

3. 设 $V=\begin{bmatrix} V_1 \\ V_2 \end{bmatrix}\in \mathbb{C}^{n\times l}(n\geq l)$满足 $V^*V=I$。证明 V_1 的奇异值皆小于等于 1。

4. 设 $A\in \mathbb{C}_p^{m\times n}$。试求出 $A_k\in \mathbb{C}_k^{m\times n}(1\leq k\leq p)$，使得

$$\|A-A_k\|_F=\min\{\|A-B\|_F:\ B\in \mathbb{C}_k^{m\times n}\};$$

如果进一步假定 A 的奇异值为

$$\sigma_1\geq\cdots\geq\sigma_n,$$

试证：

$$\min_{\operatorname{rank}(B)\leq k}\|A-B\|_F=(\sigma_{k+1}^2+\cdots+\sigma_n^2)^{\frac{1}{2}}.$$

5. 设 $\mathscr{X}$ 和 $\mathscr{Y}$ 是 $\mathbb{C}^n$ 中两个同维数的线性子空间。证明：

$$\operatorname{dist}(\mathscr{X},\mathscr{Y})=\|(I-P_{\mathscr{X}})P_{\mathscr{Y}}\|_2=\|(I-P_{\mathscr{Y}})P_{\mathscr{X}}\|_2.$$

6. 设 $A=[a_{ij}]\in \mathbb{C}^{n\times n}$。证明：

$$|\det(A)|^2\leq\min\left\{\prod_{j=1}^{n}\left(\sum_{i=1}^{n}|a_{ij}|^2\right),\ \prod_{i=1}^{n}\left(\sum_{j=1}^{n}|a_{ij}|^2\right)\right\}.$$

7. 设 $A\in \mathbb{C}^{m\times n}$ 有奇异值 $\sigma_1\geq\cdots\geq\sigma_n\geq 0$。证明：

$$\sigma_i=\max_{\substack{\mathscr{X}\\ \dim\mathscr{X}=i}}\ \min_{\substack{x\in\mathscr{X}\\ x\neq 0}}\frac{\|Ax\|_2}{\|x\|_2}=\min_{\substack{\mathscr{X}\\ \dim\mathscr{X}=n-i+1}}\ \max_{\substack{x\in\mathscr{X}\\ x\neq 0}}\frac{\|Ax\|_2}{\|x\|_2};$$

特别有

$$\sigma_1=\max_{x\neq 0}\frac{\|Ax\|_2}{\|x\|_2},\qquad \sigma_n=\min_{x\neq 0}\frac{\|Ax\|_2}{\|x\|_2}.$$

8. 设 $\alpha_1\leq\cdots\leq\alpha_n$，$\beta_1\leq\beta_2\leq\cdots\leq\beta_n$。证明

$$\min_{\pi\in\mathscr{S}_n}\sum_{i=1}^{n}(\alpha_i-\beta_{\pi(i)})^2=\sum_{i=1}^{n}(\alpha_i-\beta_i)^2.$$

9. 设 $A\in \mathbb{C}^{n\times n}$，$\lambda(A)=\{\lambda_i\}$，$\sigma(A)=\{\sigma_i\}$。试证：如果

$$|\lambda_1|\geq\cdots\geq|\lambda_n|,\ \sigma_1\geq\cdots\geq\sigma_n,$$

则

$$\prod_{i=1}^{k}\sigma_i \geqslant \prod_{i=1}^{k}|\lambda_i|, \quad k=1,2,\cdots,n.$$

10．设 $A\in\mathbb{C}^{n\times m}$，$B\in\mathbb{C}^{m\times n}$．试证 $|AB|\leqslant|A||B|$．

11．设 $A\in\mathbb{R}^{n\times n}$ 是非负不可分矩阵．证明 A 的特征多项式具有如下形式：

$$p(t)=t^m(t^h-\rho(A)^h)(t^h-\delta_2\rho(A)^h)\cdots(t^h-\delta_r\rho(A)^h),$$

其中 $m+rh=n$，$0<|\delta_i|<1$，$i=2,\cdots,r$．

第二章 矩阵计算概论

§1 矩阵计算的基本问题和来源

1.1 基本问题

矩阵计算的三类基本问题是：

（1）求解线性方程组，即给定 n 阶非奇异方阵 A 和 n 维向量 b，求一个 n 维向量 x 使得

$$Ax = b;$$

（2）求超定方程组的最小二乘解，即给定 $m\times n$ 矩阵 $A(m\geqslant n)$ 和 m 维向量 b，求 n 维向量 x 使得

$$\|Ax-b\|_2=\min\{\|Av-b\|_2:\ v\in\mathbb{R}^n\};$$

（3）计算一个矩阵的特征值和特征向量，即给定一个方阵 A，求它的全部或部分特征值，或者相应的特征向量。

这三类基本问题的来源是极其丰富的，尤其在近似求解物理中的线性偏微方程时，最终都要归结为一个矩阵计算问题。下面我们给出三个较典型的例子，以便读者对这些问题的实际背景及一个实际问题如何转化为一个矩阵计算问题有一个基本的了解。

1.2 膜的振动

现在我们考虑一个物理问题：一张弹性膜张在一个刚性的框架上，膜上每一点都受一个垂直于膜的力的作用，决定膜上每一点的垂直偏移。

在某些简化假设之下，这一物理问题可归结为求一个函数 u：$\bar{\Omega}\longrightarrow\mathbb{R}$ 满足

$$
\begin{cases}
-\Delta u(x)=f(x), & x\in\Omega, \quad (1.1)\\
u(x)=g(x), & x\in\Gamma, \quad (1.2)
\end{cases}
$$

其中Ω是平面$\mathbb{R}^2$上的一个有界连通开集，$\bar{\Omega}$表示Ω的闭包，Γ表示Ω的边界，$\Delta=\frac{\partial^2}{\partial x_1^2}+\frac{\partial^2}{\partial x_2^2}$为 Laplace 算子，$f$和$g$是已知函数。

在对f，g及边界Γ作某些适当的正则性假设之下，可以证明由(1.1)和(1.2)确定的边值问题有且仅有唯一的解，它在$\bar{\Omega}$上连续，在Ω上二次连续可微。然而，除了某些很罕见的特殊情形之外，我们一般无法求出它的精确解。因而，就需要寻找一种求其近似解的方法。有限差分法就是为求这类问题的近似解而建立起来的一种方法。

为了下面的叙述简单起见，我们假定Ω是边长为1的单位正方形，即

$$
\Omega=\{x=(x_1,x_2)\in\mathbb{R}^2\,|\,0<x_1,x_2<1\}.
$$

所谓有限差分法就是在开集Ω的有限个点上给出解的近似值，其具体做法可分为三步：

第一步　将正方形Ω以步长$h=\frac{1}{n}$分为若干正方形网格，其结点为(ih,jh)，$i,j=0,1,\cdots,n$。

第二步　在每个内部结点p（即$p\notin\Gamma$）上以有限差商代替偏导数，就可导出 Laplace 算子的五点逼近公式：

$$
\Delta_h u(p)=\frac{1}{h^2}(u(p_1)+u(p_2)+u(p_3)+u(p_4)-4u(p)),
$$

其中p_1,p_2,p_3和p_4是与p最邻近的四个结点（如图 1.1 所示）。然后在(1.1)中用$\Delta_h u(p)$代替$\Delta u(p)$，边值问题(1.1)和(1.2)就离散成

$$
\begin{cases}
-\Delta_h u(p)=f(p), & p=(ih,jh)\in\Omega,\\
u(p)=g(p), & p=(ih,jh)\in\Gamma.
\end{cases}
$$

再对结点进行编号，上面离散化方程就可写成如下形式的线性方程组

$$A_h u_h = b_h, \tag{1.3}$$

其中 A_h 为一个每行至多有五个非零元素的$(n-1)^2$阶方阵，u_h 就是内部结点上待求的 $u(p)$的值按选定顺序排列成的$(n-1)^2$维未知向量，b_h 为由 f 和 g 确定的已知向量。

第三步　解线性方程组(1.3)。

这样就把求膜的振动规律问题转化为求解线性方程组的问题。

现在假定 $h=\frac{1}{4}$，我们来考察一下几种常见的排序方法所产生的 A_h 的具体形式。

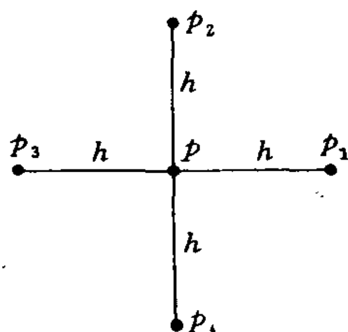


图　1.1

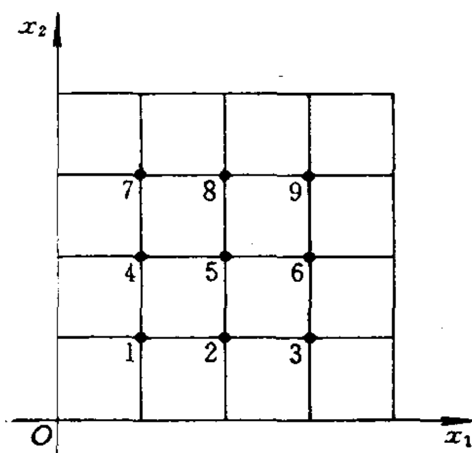


图　1.2

(1) 自然排序法。将内部结点按如图 1.2 所示的自然次序编号。此时，容易导出

$$A_h = \begin{bmatrix} B & -I & 0 \\ -I & B & -I \\ 0 & -I & B \end{bmatrix},$$

其中

$$B=\begin{bmatrix}4&-1&0\\-1&4&-1\\0&-1&4\end{bmatrix},$$

I 是三阶单位阵.

（2）奇偶排序法. 将内部节点按图 1.3 所示的次序编号，即先从小到大排 $i+j$ 为偶数的节点，然后再从小到大排 $i+j$ 为奇数的节点 (ih,jh). 在如此规定的顺序之下，有

$$A_h=\begin{bmatrix}D_1&C\\C^{\mathrm{T}}&D_2\end{bmatrix},$$

其中 $D_1=\mathrm{diag}(4,4,4,4,4)$，$D_2=\mathrm{diag}(4,4,4,4)$，

$$C=\begin{bmatrix}-1&-1&0&0\\-1&0&-1&0\\-1&-1&-1&-1\\0&-1&0&-1\\0&0&-1&-1\end{bmatrix}.$$

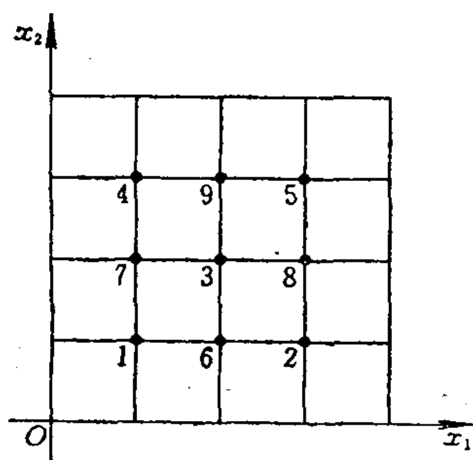


图 1.3

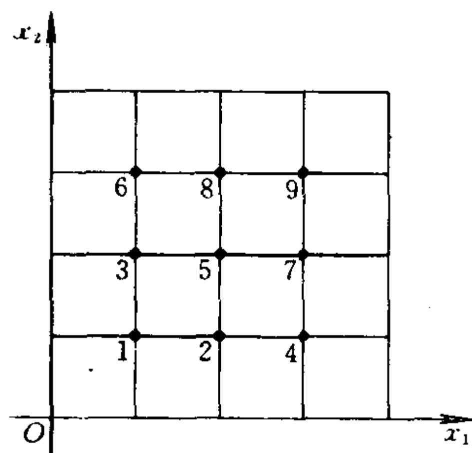


图 1.4

（3）对角排序法. 将内部节点按图 1.4 所示的对角线次序编号. 此时，系数矩阵 A_h 为

$$A_h=\begin{bmatrix}A_1&B_1&0\\B_1^{\mathrm{T}}&D&B_2\\0&B_2^{\mathrm{T}}&A_2\end{bmatrix},$$

其中

$$A_1=\begin{bmatrix}4 & -1 & -1\\ -1 & 4 & 0\\ -1 & 0 & 4\end{bmatrix},\quad D=\begin{bmatrix}4 & 0 & 0\\ 0 & 4 & 0\\ 0 & 0 & 4\end{bmatrix},$$

$$B_1=\begin{bmatrix}0 & 0 & 0\\ -1 & -1 & 0\\ 0 & -1 & -1\end{bmatrix},\quad B_2=\begin{bmatrix}-1 & 0 & 0\\ -1 & -1 & 0\\ 0 & -1 & 0\end{bmatrix},$$

$$A_2=\begin{bmatrix}4 & 0 & -1\\ 0 & 4 & -1\\ -1 & -1 & 4\end{bmatrix}.$$

由此可见，不同的排序所产生的矩阵 A_h 大不相同，因而排序将会对离散后的线性方程组的求解有极大的影响；然而，不论怎样排序，所得到的 A_h 每行总是至多有五个非零元素。因此当 n 较大时，A_h 是一个大型稀疏矩阵。例如，当 $h=1/21$ 时，A_h 是 400×400 方阵，而其中仅有大约 2000 个非零元素。

1.3 弹性系统的振动

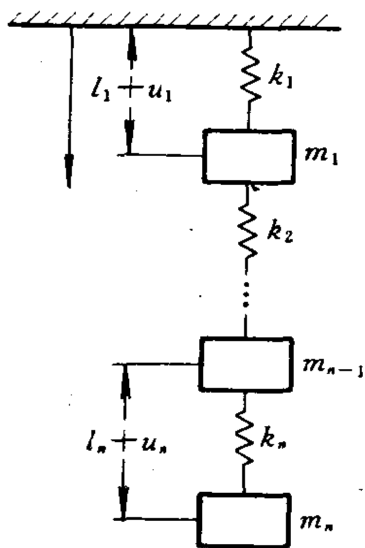


图 1.5

特征值问题的一个丰富的来源就是对振动问题的研究。下面考虑一个如图 1.5 所示的弹性系统的振动模型。它由垂直悬挂的 n 段弹簧和 n 个物体构成，在重力和弹性力的作用下，在垂直方向作振动。我们的问题是求系统中每个物体的位移。

设第 i 个物体的质量为 m_i，第 i 段弹簧的弹性系数是 k_i；并假定在时刻 t 第 i 个物体的位移为 $u_i(t)$。按照力学的有关理论可导出系统的运动方程是

$$\ddot{u} = Au, \tag{1.4}$$

其中 $u=(u_1,\cdots,u_n)^{\mathrm{T}}$，$\ddot{u}=\left(\frac{\mathrm{d}^2u_1}{\mathrm{d}t^2},\cdots,\frac{\mathrm{d}^2u_n}{\mathrm{d}t^2}\right)^{\mathrm{T}}$，

$$A=\begin{bmatrix} \alpha_1 & \gamma_1 & & & 0 \\ \beta_1 & \alpha_2 & \gamma_2 & & \\ & \beta_2 & \ddots & \ddots & \\ & & \ddots & \ddots & \gamma_{n-1} \\ 0 & & & \beta_{n-1} & \alpha_n \end{bmatrix},$$

$$\alpha_i = -\left(\frac{k_{i+1}}{m_i}+\frac{k_i}{m_i}\right) \quad (k_{n+1}=0),$$

$$\beta_i = k_{i+1}/m_{i+1},$$

$$\gamma_i = k_{i+1}/m_i.$$

如果我们试求(1.4)的形如

$$u = u_0 \sin\lambda t$$

的解，其中 $u_0=(u_1^{(0)},\cdots,u_n^{(0)})^{\mathrm{T}}\in\mathbb{R}^n$ 为不依赖于 t 的常向量，则代入(1.4)有

$$(-\lambda^2\sin\lambda t)u_0=(\sin\lambda t)Au_0,$$

即 $-\lambda^2$ 和 u_0 应满足

$$Au_0=-\lambda^2u_0.$$

这样，要求系统中每个物体位移的问题就转化成求 A 的特征值 $-\lambda^2$ 和特征向量 u_0 的问题，这里的 λ 在物理上叫做系统的**固有频率**。

1.4 多元线性回归分析

在现实生活中，我们常常希望根据以往的经验对未来可能发生的事件作出预测。例如，在气象中，希望基于在某几个时刻对大气压的测量，预测几小时以后的气象情况。这类问题用数学的语言来讲，就是要求研究某个随机变量 Y 与另外一些随机变量

$X_1,\cdots,X_m$ 之间的关系，并通过对 $X_1,\cdots,X_m$ 的测量，预测在今后可以观察到的变量 Y。回归分析方法就是处理这类问题的一种常用的方法。这里，我们来考察一种较简单的回归分析模型——多元线性回归分析。

设随机变量 Y 和 $X_1,\cdots,X_m$ 之间有如下的线性关系：

$$Y = b_0 + b_1X_1 + \cdots + b_mX_m,$$

但其中的 $b_0,b_1,\cdots,b_m$ 是未知常数；并设对这些变量进行了 n 次观察得到如下数据：

$$y_j;x_{1j},\cdots,x_{mj},\quad j=1,2,\cdots,n.$$

线性回归分析就是根据这 n 次观察到的数据对 $b_0,\cdots,b_m$ 给出估计。通常采用的方法是求 $\hat{b}_0,\cdots,\hat{b}_m$ 使得

$$\sum_{i=1}^{n}\left(y_i-\left(\sum_{j=1}^{m}x_{ji}\hat{b}_j+\hat{b}_0\right)\right)^2$$

$$=\min\left\{\sum_{i=1}^{n}\left(y_i-\left(\sum_{j=1}^{m}x_{ji}b_j+b_0\right)\right)^2:\ b_j\in\mathbb{R},\ j=0,1,\cdots,m\right\}.$$

将其写成矩阵向量形式就是求最小二乘问题

$$\|Xb-y\|_2=\min\{\|Xv-y\|_2:\ v\in\mathbb{R}^{m+1}\}$$

之解，其中 $b=(b_0,b_1,\cdots,b_m)^{\mathrm{T}}$，$y=(y_1,\cdots,y_n)^{\mathrm{T}}$，

$$X=\begin{bmatrix}1 & x_{11} & x_{21} & \cdots & x_{m1}\\ 1 & x_{12} & x_{22} & \cdots & x_{m2}\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ 1 & x_{1n} & x_{2n} & \cdots & x_{mn}\end{bmatrix}.$$

这样一来，一个需预测的问题，就通过多元线性回归分析的方法，最终归结为一个线性最小二乘问题的求解问题。

§2 病态问题和数值稳定性

这一节，我们着重介绍矩阵计算中两个非常重要的基本概念，即矩阵计算问题的病态性和计算方法的数值稳定性。

2.1 矩阵计算问题的病态和良态

对于一个给定的矩阵计算问题，由于误差的存在，我们首先必须研究的一个重要问题是：问题的参数（如在线性方程组的求解问题中，系数矩阵和右端项即为参数）的微小扰动对问题的解将会产生什么样的影响呢？这就是所谓的问题的解对参数扰动的**敏感性**问题。不同的问题其解对参数的扰动的敏感程度大不相同，有的十分敏感，有的则不然。如果一个矩阵计算问题，其参数的微小变化会引起解的巨大变化，则称这个矩阵计算问题本身是**病态的**；否则称其是**良态的**。病态的矩阵计算问题是经常遇到的，现在来看一个简单的例子。

例2.1 求解如下的线性方程组

$$\begin{bmatrix} 1.001 & 0.999 \\ 0.999 & 1.001 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}. \tag{2.1}$$

容易看出，它的解为 $x=(x_1,x_2)^T=(1,1)^T$。

如果把(2.1)的右端项作微小扰动

$$\delta b=(10^{-3},\ -10^{-3})^T,$$

则(2.1)变为

$$\begin{bmatrix} 1.001 & 0.999 \\ 0.999 & 1.001 \end{bmatrix}\begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix} = \begin{bmatrix} 2.001 \\ 1.999 \end{bmatrix}. \tag{2.2}$$

直接计算可知，(2.2)的解为 $\tilde{x}=(\tilde{x}_1,\tilde{x}_2)^T=(1.5,0.5)^T$。因此有

$$\frac{\|\tilde{x}-x\|_\infty}{\|x\|_\infty}=0.5,\quad \frac{\|\delta b\|_\infty}{\|b\|_\infty}=\frac{1}{2000}.$$

这表明解的相对误差是右端项相对误差的1000倍！

如果把(2.1)的系数矩阵作微小扰动变为

$$\begin{bmatrix} 1 & 1 \\ 0.999 & 1 \end{bmatrix}\begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix}=\begin{bmatrix} 2 \\ 2 \end{bmatrix},$$

则其解变为$\tilde{x}=(0,2)^{\mathrm{T}}$，这和(2.1)的解的差别就更大了。

这表明线性方程组(2.1)的解对参数的微小扰动是十分敏感的。

这里需要特别强调的一点是：一个矩阵计算问题的病态与否，是这个问题本身的固有属性，与计算这个问题的算法无关。当然，问题的病态性会对它的近似求解带来很大的困难，病态程度越严重，困难就越大。

此外，病态与良态之间并没有严格的界限，是相对而言的，是由所考虑的问题的实际背景和所使用的计算机的精度等因素来确定的。

上面所讲的病态或良态只是一种定性的描述，而从实际计算的角度出发，还需从定量上加以刻画。有关三类基本矩阵计算问题的解对扰动的敏感性的定量刻画将在今后的有关章节里详细加以介绍。

2.2 算法的数值稳定性

所谓计算一个数学问题的算法就是一些数据按照指定的顺序进行运算的一个序列。例如，对于数学问题：已知$p^2>4q$，求

$$x=\frac{1}{2}(-p+\sqrt{p^2-4q})$$

的值可设计如下两个算法：

算法 I	算法 II
(1) $l=4q$,	(1)—(4) 同算法 I，
(2) $s=p^2$,	(5) $y=r+p$,
(3) $t=s-l$,	(6) $z=2q$,
(4) $r=\sqrt{t}$,	(7) $x=-z/y$,

(5) $y=r-p$,

(6) $x=\frac{1}{2}y$,

其中算法Ⅱ是根据 x 的等价表示 $x=-2q/(p+\sqrt{p^2-4q})$ 设计的。

一个好的算法通常应该具有如下特征:

(1) 运算量小;

(2) 舍入误差对计算结果的影响小;

(3) 计算过程的中间结果占用存储空间少;

(4) 易于程序实现.

也就是说，一个好的算法应该既快又准。但要设计一个好的算法是非常困难的。有时，上述几条并非能够同时满足，这就需根据具体问题有所偏重，有的偏重于快，有的偏重于准。

这里需要特别指出的是，在设计算法时，必须充分注意舍入误差对计算结果的影响。对于同一个计算问题，不同的算法，由于舍入误差的积累不同，其计算结果可能大不相同。即使一个非常良态的问题，由于使用的方法不当，也可使计算结果完全失真，而变得毫无用处。请看下面一个简单的例子。

例2.2 考虑线性方程组

$$\begin{cases} 10^{-11}x_1+x_2=1, \\ (1+10^{-11})x_1+x_2=2. \end{cases} \tag{2.3}$$

容易验证这是一个十分良态的矩阵计算问题。

现在假定我们是在十位十进制浮点数系下求解这一方程组。若按自然顺序由第二个方程减去第一个方程两边乘以 10^{11} 消去 x_1 来求解，则得 $x_1=0$，$x_2=1$；而若用第一个方程减去第二个方程消去 x_2 来求解，则得 $x_1=1$，$x_2=1$。与(2.3)的精确解 $x_1=1$，$x_2=1-10^{-11}$ 来比可知，前者已面貌全非，而后者却已精确到小数点之后第十位，已是能够得到的最好结果。

上例说明，对于同一计算问题，使用的方法不同，效果也大不相同。一个算法；如果计算过程的舍入误差的积累不大，就说这一算法具有**较好的数值稳定性**；反之，如果计算过程的舍入误差的积累很大，就说该算法**数值稳定性不好**，或者说其是**数值不稳定的**。

要想判定一个算法的数值稳定性的好坏，需要应用舍入误差分析的方法。关于舍入误差分析，通常有两种分析方法。一种是**向前误差分析法**，是根据浮点运算的舍入误差规律，直接估计计算结果和真实结果之间的误差。例如，对于加法运算，要估计

$$|fl(x+y)-(x+y)|$$

的上界，其中 $fl(x+y)$ 表示 $x+y$ 在浮点数系下的计算结果。另一种是**向后误差分析法**，是根据浮点运算的舍入误差规律，把计算过程的误差返回到原始数据的误差。例如，对于加法运算，首先将 x 和 y 的实际浮点运算表为

$$fl(x+y)=x(1+\delta)+y(1+\delta),$$

即看作是对 $x(1+\delta)$ 和 $y(1+\delta)$ 的精确加法运算；然后再对 δ 的大小给出估计。由于向后误差分析法将计算机中浮点数的实际计算转化为通常的实数的精确运算，所以在分析过程中就可毫无困难地使用实数运算的代数运算法则，而向前误差分析就没有这一优点。因此在实际分析时，经常使用的是向后误差分析法。

现在，我们以线性方程组的求解为例说明如何利用向后误差分析法来判定一个算法的数值稳定性。设应用某种算法（例如，Gauss 消去法）求解线性方程组 $Ax=b$ 得到的计算解为 $\tilde{x}$。利用向后误差分析方法，可将 $\tilde{x}$ 归结为扰动方程组

$$(A+\delta A)\tilde{x}=b+\delta b$$

的精确解，并给出 $\|\delta A\|$ 和 $\|\delta b\|$ 的上界估计。在这里起主导作用的是 $\|\delta A\|$ 的大小，不同的算法所产生的 δA，其大小各不相同。

因此，我们用$\|\delta A\|/\|A\|$的大小来衡量算法的数值稳定性的好坏，它越小，表示该算法的数值稳定性越好。

最后需指出的一点是，算法的数值稳定性的好坏与问题的病态性一样，也是就相互对比而言的，是算法本身所固有的属性，与所要计算的具体问题是否病态无关。

§3 矩阵计算的基本工具

矩阵计算的基本途径就是设法把一个较复杂的矩阵计算问题转化为一个简单的、易于求解的矩阵计算问题，而完成这一转化过程所使用的主要"工具"有三种：Householder 变换，Givens 变换和 Gauss 变换。这一节，我们就来介绍这三种变换。

3.1 Householder 变换

Householder 变换是形如

$$H = I - 2ww^{\mathrm{T}} \tag{3.1}$$

的 n 阶实方阵，其中 $w \in \mathbb{R}^n$，且$\|w\|_2 = 1$；有时亦称这样的矩阵为**Householder 矩阵或镜像变换**。Householder 变换有许多良好的性质，下面的定理列举了几条最基本的性质。

定理3.1 设H是如(3.1)所定义的 Householder 变换，则

(1) H是实对称的正交矩阵；

(2) H仅有两个互不相同的特征值 -1 和 1，其中 1 是 $n-1$ 重的，-1 是单重的，而且 w 就是属于 -1 的单位特征向量；

(3) $\det(H) = -1$；

(4) 对任意的 $x \in (\mathrm{span}\{w\})^{\perp}$ 和 $\alpha \in \mathbb{R}$，有

$$H(x + \alpha w) = x - \alpha w,$$

即H是关于w的垂直超平面的反射变换(如图 3.1 所示)。

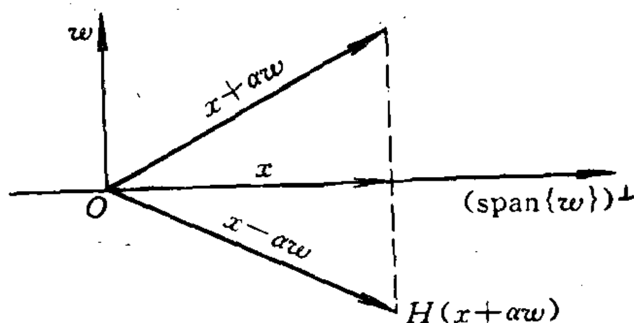


图 3.1

证明留作练习。

Householder 变换之所以在矩阵计算中占有重要的地位，是因为它具有如下定理所述的性质。

定理3.2 设 $x\in\mathbb{R}^n$ 是任一给定的非零向量. 则可构造出单位向量 $w\in\mathbb{R}^n$，使按(3.1)定义的 Householder 变换 H 满足

$$Hx=\alpha e_1, \tag{3.2}$$

其中 $\alpha=\pm\|x\|_2$。

证明 由于

$$Hx=(I-2ww^{\mathrm{T}})x=x-2(w^{\mathrm{T}}x)w,$$

所以欲使(3.2)成立，必须有

$$2(w^{\mathrm{T}}x)w=x\pm\|x\|_2e_1.$$

由此立知，w 应为

$$w=\frac{x\pm\|x\|_2e_1}{\|x\pm\|x\|_2e_1\|_2}. \tag{3.3}$$

当然，这里假定 $x\pm\|x\|_2e_1\neq0$. 直接验证可知这样定义的w满足定理的要求。证毕。

定理 3.2 告诉我们，对于任意的 $x\in\mathbb{R}^n$，$x\neq0$，都可构造出 Householder 变换H，将 x 的后 $n-1$ 个分量变为零；而其证明又告诉我们，可按如下步骤来构造 w：

(1) 计算 $v=x\pm\|x\|_2e_1$;

(2) 计算 $w=v/\|v\|_2$。

然而，从实际计算的角度出发，仍有两个具体问题需要解决：

(1) 计算 v 时，$\|x\|_2$ 前的符号如何选取最好？

(2) x 分量模很大时，如何避免计算 $\|x\|_2$ 时可能出现的溢出？

首先考虑 $\|x\|_2$ 前的符号选取问题。如果 x 是一个很接近于 e_1 的向量，则 $v=x-\text{sign}(\xi_1)\|x\|_2 e_1$ 就很接近于 0（其中 ξ_1 表示 x 的第一个分量），从而单位化时就会产生较大的误差。因此，为了避免这种现象出现，我们应选取 $\|x\|_2$ 前的符号与 x 的第一个分量的符号相同，即应取

$$v=x+\text{sign}(\xi_1)\|x\|_2 e_1. \tag{3.4}$$

这保证了 $\|v\|_2\geqslant\|x\|_2$，且误差分析表明，这样选择的 v，可使实际计算得到的 H 具有良好的正交性。

再来考虑如何避免计算 $\|x\|_2$ 可能产生的溢出问题。由于对任意的非零常数 α，αv 与 v 的单位化向量是一样的，因此我们可以用 $x/\|x\|_\infty$ 代替 x 来构造 v（这相当于在原来的 v 之前乘了因子 $\alpha=1/\|x\|_\infty$）。

另外还需指出的是，实际计算时，也不需明确地将 v 单位化。由于有

$$H=I-2ww^{\mathrm{T}}=I-2\frac{vv^{\mathrm{T}}}{\|v\|_2^2}=I-\beta vv^{\mathrm{T}},$$

其中 $\beta=2/v^{\mathrm{T}}v$，因此，实际上只需将 v 和 β 求出即可。

综上所述可得如下的基本算法。

算法3.1

(1) 输入 $x=(\xi_1,\cdots,\xi_n)^{\mathrm{T}}$。

(2) $\eta:=\max\{|\xi_1|,\cdots,|\xi_n|\}$。

(3) 如果 $\eta=0$，则 $\beta:=0$，转步(7)。

(4) $\xi_i:=\xi_i/\eta\ (i=1,2,\cdots,n)$，$\alpha:=\left(\sum_{i=1}^{n}\xi_i^2\right)^{1/2}$。

(5) 如果$\xi_1<0$，则$\alpha:=-\alpha$。

(6) $\xi_1:=\xi_1+\alpha, \beta:=\frac{1}{\alpha\xi_1}, \alpha:=\eta\alpha$。

(7) 输出有关信息，结束。

这一算法对给定的向量x计算出了数β和向量$v=(v_1,\cdots,v_n)^T$，使Householder变换$H=I-\beta vv^T$满足$Hx=-\alpha$，并将v就存在x所占用的存储单元内。它需要做乘法$n+1$次，除法$n+1$次，加法$n+1$次，开方1次。

由此可见，这一算法的主要工作量是乘除运算。一般关于矩阵计算问题的算法都有类似的特点。因此，习惯上在考察一个算法的运算量时，只计算它的乘除运算的次数，而且通常忽略这一次数中关于n的低阶项不计，而只考虑它的高阶项。因此，以后凡谈到一个算法的运算量皆指这一算法乘除运算次数的高阶项。例如，某一算法的乘除运算的次数是$\frac{1}{3}n^3+50n^2+100n+200$，则我们就说这一算法的运算量为$\frac{1}{3}n^3$。按照这样的约定，算法3.1的运算量是$2n$。

在应用Householder变换约化矩阵时，主要工作就是要计算一个矩阵和一个Householder矩阵的乘积。在实际计算时，我们不是将Householder矩阵明确地计算出来，然后再作两个矩阵的乘积，而是充分利用Householder矩阵的特殊结构来进行计算。

设$A\in\mathbb{R}^{n\times q}$，$H=I-\beta vv^T$。则

$$HA=(I-\beta vv^T)A=A-\beta v(A^Tv)^T. \tag{3.5}$$

因此，我们只要知道构成H的向量v和常数β，就可按照(3.5)来计算HA。这样，假定$v=(0,\cdots,0,v_k,\cdots,v_n)^T\in\mathbb{R}^n$，$\beta=2/v^Tv$，并假定$HA$的计算结果就存储在$A$所占用的存储单元里，就可得如下的基本算法。

算法3.2

(1) 输入 A, v 和 β，$j := 1$。

(2) $\sigma := \nu_k \alpha_{kj} + \cdots + \nu_n \alpha_{nj}$,

$\sigma := \beta\sigma$,

$\alpha_{ij} := \alpha_{ij} - \sigma\nu_i \quad (i = k, \cdots, n)$。

(3) 如果 $j < q$，则 $j := j+1$，转 步(2)；否则 输出有关信息，结束。

容易算出，这一算法的运算量是 $2q(n-k+1)$。

算法 3.1 和 3.2 的数值性态是十分令人满意的。假定算法3.1的计算结果是 $\hat{v}$ 和 $\hat{\beta}$。定义

$$\hat{H} = I - \hat{\beta}\hat{v}\hat{v}^{\mathrm{T}},$$

则可证（参见文献[71]）

$$\|H - \hat{H}\| \leqslant 10\varepsilon,$$

其中 $H = I - \beta vv^{\mathrm{T}}$ 是准确的 Householder 矩阵，ε 是机器精度（亦称单位误差），即

$$\varepsilon = \begin{cases} b^{1-t}/2, & \text{当舍入法被使用时}, \\ b^{1-t}, & \text{当截断法被使用时}, \end{cases}$$

其中 b 是机器所用浮点数的基底，t 是字长。

假定 $\hat{\beta}$ 和 $\hat{v}$ 又被用在算法 3.2 中去计算 $\hat{H}A$，并假定计算结果是 $\hat{A}$，则

$$\hat{A} = H(A+E),$$

其中
$$\|E\|_2 \leqslant c(n-k+1)^2 \|A\|_2 \varepsilon,$$

此处 c 是一个不依赖于 n 的常数。这就是说，计算所得到的 Householder 矩阵 $\hat{H}$，再作用在一个矩阵 A 上，就相当于准确的 H 作用在一个与 A 十分靠近的矩阵 $A+E$ 上。

3.2 Givens 变换

Givens变换就是形如

$$
G(i,k,\theta)=\begin{bmatrix}
1 & & & & & & & & \\
& \ddots & & & & & & & \\
& & 1 & & & & & & \\
& & & c & \cdots & s & & & \\
& & & \vdots & 1 \ \ddots & \vdots & & & \\
& & & & 1 & & & & \\
& & & -s & \cdots & c & & & \\
& & & & & & 1 & & \\
& & & & & & & \ddots & \\
& & & & & & & & 1
\end{bmatrix}
\begin{matrix} \\ \\ \\ i \\ \\ \\ k \\ \\ \\ \\ \end{matrix}
\tag{3.6}
$$

$\qquad\qquad\qquad\qquad i \qquad\qquad\qquad k$

的$n\times n$矩阵，其中$c=\cos\theta$，$s=\sin\theta$，θ是某一实数。

$G(i,k,\theta)$是一个正交矩阵，而且用它作用在一个向量上，就相当于在(i,k)坐标平面内作一个角度为θ的平面旋转．因此，有时亦称Givens变换为**平面旋转变换**。

设$x=(\xi_1,\cdots,\xi_n)^{\mathrm{T}}\in\mathbb{R}^n$．令$y=(\zeta_1,\cdots,\zeta_n)^{\mathrm{T}}=G(i,k,\theta)x$，则

$$
\begin{aligned}
\zeta_i &= c\xi_i+s\xi_k,\\
\zeta_k &= -s\xi_i+c\xi_k,\\
\zeta_j &= \xi_j,\quad j\neq i \text{ 或 } k.
\end{aligned}
$$

从这些公式可以看出，如果我们希望变换后的向量y的第k个分量为零，则只需取

$$
c=\frac{\xi_i}{\sqrt{\xi_i^2+\xi_k^2}},\qquad s=\frac{\xi_k}{\sqrt{\xi_i^2+\xi_k^2}} \tag{3.7}
$$

即可，而实际计算时，为了尽可能地减少运算次数，避免可能出现的溢出现象，并考虑到算法的数值稳定性，对给定的$x=(\xi_1,$

$\cdots,\xi_n)^T\in\mathbb{R}^n$，将按如下的基本算法来计算 $c=\cos\theta$，$s=\sin\theta$使 $G(i,k,\theta)x$ 的第 k 个分量为零。

算法3.3

(1) 输入ξ_k和ξ_i。

(2) 如果 $\xi_k=0$，则 $c:=1$，$s:=0$，转步(4)；否则进行下一步。

(3) 如果$|\xi_k|\geqslant|\xi_i|$，则

$$t:=\xi_i/\xi_k,\quad s:=1/(1+t^2)^{1/2},\quad c:=st;$$

否则

$$t:=\xi_k/\xi_i,\quad c:=1/(1+t^2)^{1/2},\quad s:=ct.$$

(4) 输出 c 和 s，结束。

在实际应用时，经常需要计算一个Givens变换与一个矩阵的乘积。设$A\in\mathbb{R}^{n\times q}$。则$G(i,k,\theta)A$只改变矩阵$A$的第 i 行和第 k 行元素之值，其余元素不变．利用这一性质，如果我们已知 $G(i,k,\theta)$的指标 i 和 k，以及$c=\cos\theta$与 $s=\sin\theta$，并将 $G(i,k,\theta)A$ 的计算结果就存放在A所用的存储单元内，则可得如下算法：

算法3.4

(1) 输入 A,i,k,c,s；$j:=1$。

(2) $\alpha:=a_{ij}$，$\beta:=a_{kj}$，

$$a_{ij}:=c\alpha+s\beta,\qquad a_{kj}:=-s\alpha+c\beta.$$

(3) 如果 $j<q$，则$j:=j+1$，转步(2)；否则输出有关信息，结束。

这一算法的运算量是$4q$。

Givens 变换的数值性态亦是良好的。假定$\hat{c}$和$\hat{s}$是由算法(3.3)产生的，则

$$\hat{c}=c(1+\varepsilon_c),\qquad \varepsilon_c=O(\varepsilon),$$
$$\hat{s}=s(1+\varepsilon_s),\qquad \varepsilon_s=O(\varepsilon).$$

如果 $\hat{c}$ 和 $\hat{s}$ 又被用在算法3.4中，并记计算结果为 $\tilde{A}$，则有

$$\tilde{A}=G(i,k,\theta)(A+E),$$

其中

$$\|E\|_2/\|A\|_2=O(\varepsilon).$$

详细的误差分析参见文献[71]

3.3 Gauss 变换

前面介绍的两种变换都是正交变换，而且它们都有把一个向量的若干指定位置的分量变为零的功能。这里我们再介绍一种具有同样功能的非正交变换，即所谓的 Gauss 变换。

设 $x=(\xi_1,\cdots,\xi_n)^{\mathrm T}\in\mathbb{R}^n$，满足 $\xi_k\neq 0$。令

$$l_{ik}=\xi_i/\xi_k,\qquad i=k+1,\cdots,n, \tag{3.8}$$

并定义

$$L_k=I-l_ke_k^{\mathrm T}, \tag{3.9}$$

其中 $l_k=(\underbrace{0,\cdots,0}_{k},l_{k+1,k},\cdots,l_{nk})^{\mathrm T}$，则有

$$L_kx=x-\xi_kl_k=(\xi_1,\cdots,\xi_k,0,\cdots,0)^{\mathrm T}.$$

形如(3.9)的矩阵 L_k 被称作**Gauss变换**，有时亦被称作**初等下三角阵**，其中 l_k 被称作**Gauss 向量**。

用Gauss 变换作用在一个向量上，其运算特别简单，比如说 $y=(\zeta_1,\cdots,\zeta_n)^{\mathrm T}\in\mathbb{R}^n$，则 $L_ky=y-\zeta_kl_k$。利用这一性质，可给出计算 L_kA 的算法如下。

算法3.5

(1) 输入 $A=[a_{ij}]\in\mathbb{R}^{n\times q}$, $l_k=(0,\cdots,0,l_{k+1,k},\cdots,l_{nk})^{\mathrm T}$; $j:=1$。

(2) $a_{ij}:=a_{ij}-a_{kj}l_{ik}(i=k+1,\cdots,n)$。

(3) 如果 $j<q$，则 $j:=j+1$，转步(2)；否则输出有关信息，

结束。

这一算法的运算量是$(n-k)q$。

设$\hat{l}_k$是l_k按公式(3.8)进行计算所得到的结果，则有

$$\hat{l}_k = l_k + e, \qquad |e| \leqslant \boldsymbol{\varepsilon} |l_k|.$$

如果$\hat{l}_k$又被用在算法3.5中，并记计算结果为$\hat{A}$，则有

$$\hat{A} = (I - l_k e_k^{\mathrm{T}})(A + E),$$

其中

$$|E| \leqslant 3\boldsymbol{\varepsilon}[|A| + |l_k||a_k^{\mathrm{T}}|] + O(\boldsymbol{\varepsilon}^2),$$
$$a_k^{\mathrm{T}} = (a_{k,k+1}, \cdots, a_{kq}).$$

详细误差分析参见文献[71]。

由此可见，假如$\|l_k\|$很大，则引起的相对误差$\|E\|_\infty / \|A\|_\infty$也很大。因此，使用Gauss变换时要特别小心。

最后我们需指出的是Gauss变换的乘积所具有的一个良好性质。设

$$L = L_{n-1} L_{n-2} \cdots L_1,$$

其中

$$L_i = I - l_i e_i^{\mathrm{T}}, \qquad l_i = (0, \cdots, 0, l_{i+1,i}, \cdots, l_{ni})^{\mathrm{T}},$$

则易证L是一个单位下三角阵，而且

$$\begin{aligned} L^{-1} &= L_1^{-1} L_2^{-1} \cdots L_{n-1}^{-1} \\ &= (I + l_1 e_1^{\mathrm{T}})(I + l_2 e_2^{\mathrm{T}}) \cdots (I + l_{n-1} e_{n-1}^{\mathrm{T}}) \\ &= I + l_1 e_1^{\mathrm{T}} + \cdots + l_{n-1} e_{n-1}^{\mathrm{T}} \\ &= \begin{bmatrix} 1 & & & & \\ l_{21} & 1 & & 0 & \\ l_{31} & l_{32} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ l_{n1} & l_{n2} & \cdots & l_{n,n-1} & 1 \end{bmatrix}. \end{aligned}$$

因此，只要知道每个Gauss向量l_k，我们就可很容易将L^{-1}写出来。

习　　题

1. 证明：求解复矩阵和复向量的线性方程组可归结为求解实向量和实矩阵的线性方程组。是否亦可同样把计算复矩阵的特征值问题归结为计算实矩阵的特征值问题？

2. 证明：在作两个复数的乘积时，可以只用3次实数乘法；在作两个2阶复矩阵的乘积时，可以只用7次实数的乘法。

3. 已知

$$(\sqrt{2}-1)^6=(3-2\sqrt{2})^3=99-70\sqrt{2}$$

$$=\frac{1}{(\sqrt{2}+1)^6}=\frac{1}{(3+2\sqrt{2})^3}$$

$$=\frac{1}{99+70\sqrt{2}}.$$

请指出哪一个公式进行计算误差较小，并说明理由。

4. 举例说明在计算机上有$(a+b)c\neq ac+bc$。

5. 设计一种计算$\|x\|_2$的算法，其中$x\in\mathbb{R}^n$，并给出其舍入误差界。

6. 设x和y是$\mathbb{R}^n$中两个非零向量。给出一种算法，来确定一个Householder变换H，使得$Hx\in\operatorname{span}\{y\}$。

7. 证明：如果x和y是两个n维实向量，那么

$$\det(I+xy^{\mathrm{T}})=1+x^{\mathrm{T}}y.$$

8. 假定$x\in\mathbb{C}^2$。给出一种算法，确定一个如下形式的酉矩阵

$$Q=\begin{bmatrix} c & s \\ -\bar{s} & c \end{bmatrix},\qquad c\in\mathbb{R},\ c^2+|s|^2=1,$$

使得Qx的第二个分量为零。

9. 假定x和y是$\mathbb{R}^n$中两个单位向量。给出一种使用Givens

变换的算法，来计算一个正交矩阵 Q，使得 $Qx=y$。

10. 确定一个 3×3 的 Gauss 变换 L，使得

$$L\begin{bmatrix}2\\3\\4\end{bmatrix}=\begin{bmatrix}2\\7\\8\end{bmatrix}.$$

第三章　线性方程组的直接解法

§1　线性方程组的条件数

在第二章中我们曾对一个矩阵计算问题的病态性这个概念作了概略的定性说明。现在我们来考虑如何对线性方程组的病态程度作出定量估计。

设$A\in\mathbb{R}^{n\times n}$是非奇异的，$b\in\mathbb{R}^n$。我们来考虑线性方程组

$$Ax=b \tag{1.1}$$

之解x对数据A和b的微小扰动的敏感程度。为此，考虑如下的含参方程组

$$(A+\varepsilon\delta A)x(\varepsilon)=b+\varepsilon\delta b,\qquad x(0)=x, \tag{1.2}$$

其中$\delta A\in\mathbb{R}^{n\times n}$，$\delta b\in\mathbb{R}^n$，$\varepsilon$充分小。显然，在$\varepsilon=0$的充分小的邻域内$x(\varepsilon)=(A+\varepsilon\delta A)^{-1}(b+\varepsilon\delta b)$是$\varepsilon$的可微向量值函数，且易知

$$\dot{x}(0)=A^{-1}(\delta b-\delta Ax). \tag{1.3}$$

将(1.3)代入$x(\varepsilon)$的Taylor展式

$$x(\varepsilon)=x(0)+\varepsilon\dot{x}(0)+O(\varepsilon^2),$$

并取范数，可得

$$\frac{\|x(\varepsilon)-x\|}{\|x\|}\leqslant\varepsilon\|A^{-1}\|\left\{\frac{\|\delta b\|}{\|x\|}+\|\delta A\|\right\}+O(\varepsilon^2), \tag{1.4}$$

其中的矩阵范数是由相应的向量范数诱导出的算子范数。现定义

$$\kappa(A)=\|A\|\|A^{-1}\|, \tag{1.5}$$

并利用不等式

$$\|b\| \leqslant \|A\| \|x\|,$$

由(1.4) 可得

$$\frac{\|x(\varepsilon)-x\|}{\|x\|} \leqslant \kappa(A)\left\{\frac{\|\delta b\|}{\|b\|}\varepsilon + \frac{\|\delta A\|}{\|A\|}\varepsilon\right\} + O(\varepsilon^2). \qquad (1.6)$$

(1.6) 表明，解 x 的相对误差大约是 A 与 b 的相对误差之和的 $\kappa(A)$ 倍。从这个意义上来讲，$\kappa(A)$ 的大小反映了方程组(1.1)之解对微小扰动的敏感程度。因此，我们称 $\kappa(A)$ 为线性方程组(1.1)的求解问题的**条件数**。常用的是对应于 p 范数的条件数，通常记作 $\kappa_p(A)$；特别当 $p=2$ 时称作**谱条件数**。关于条件数的一些最基本的性质可总结为如下定理。

定理1.1 设 $A\in\mathbb{R}^{n\times n}$ 非奇异。则

(1) $\kappa(A)\geqslant 1$，$\kappa(A)=\kappa(A^{-1})$，$\kappa(A)=\kappa(\alpha A)$，$\alpha\neq 0$；

(2) $\kappa_2(A)=\sigma_1(A)/\sigma_n(A)$，其中 $\sigma_1(A)$ 和 $\sigma_n(A)$ 分别表示 A 的最大和最小奇异值；

(3) 若 A 正规，则

$$\kappa_2(A)=\max_{\lambda\in\lambda(A)}|\lambda| \Big/ \min_{\lambda\in\lambda(A)}|\lambda|;$$

(4) 若 A 是酉矩阵，则 $\kappa_2(A)=1$；

(5) $\kappa_2(A)$ 是酉不变的，即

$$\kappa_2(A)=\kappa_2(UA)=\kappa_2(AU), \quad \forall U\in\mathscr{U}_n.$$

证明留作练习。

定理1.1的(3)表明，对于一个正规矩阵而言，有“大”的谱条件数的充分必要条件是特征值的模最大者与最小者之比“大”。但这对一般方阵来讲是不成立，即使特征值都相等，它的条件数也可能很大。例如

$$A=\begin{bmatrix} 1 & 2 & & 0 \\ & 1 & \ddots & \\ & & \ddots & 2 \\ 0 & & & 1 \end{bmatrix}\in\mathbb{R}^{100\times 100} \qquad (1.7)$$

的特征值都是 1，而 $\kappa_2(A)>2^{100}$。

第二章曾考察过的例2.1，其系数矩阵 A 是对称正定的，它的最大和最小特征值分别为 $\lambda_1=2$ 和 $\lambda_2=0.002$，因此由定理 1.1的(3)知，其谱条件数为 $\kappa_2(A)=1000$。这表明其系数矩阵的条件数是很大的，因而才会出现其解对扰动十分敏感的现象。这进一步说明，条件数的大小确实在一定程度上反映了线性方程组求解问题的病态程度。

因此，当 $\kappa(A)$ 很大时，我们就说线性方程组 (1.1)的求解问题是**病态的**，或者说矩阵 A 是**病态的**；否则就说其是**良态的**。

由范数等价定理可知，对于任意两个不同范数定义的条件数 $\kappa_\alpha(\cdot)$ 和 $\kappa_\beta(\cdot)$，必存在两个正数 c_1 和 c_2，使得

$$c_1\kappa_\alpha(A)\leqslant\kappa_\beta(A)\leqslant c_2\kappa_\alpha(A),\qquad \forall A\in\mathbb{R}_n^{n\times n}.$$

因此，若一个矩阵 A 在 α 范数是病态的，即 $\kappa_\alpha(A)$ 很大，则 $\kappa_\beta(A)/c_1$ 亦很大，其中 c_1 是与 A 无关的正数；反过来，若 A 在 β 范数下有 $\kappa_\beta(A)$ 很大，则 $c_2\kappa_\alpha(A)$ 亦很大，c_2 亦是与 A 无关的正数。从这个意义上来讲，一个矩阵病态与否与具体的范数无关。

一类十分典型的病态矩阵是所谓的 Hilber 矩阵，其定义为

$$H_n=\begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{1}{n} & \frac{1}{n+1} & \frac{1}{n+2} & \cdots & \frac{1}{2n-1} \end{bmatrix}.$$

其条件数 $\kappa_2(H_n)\approx e^{3.5n}$ 随着 n 的增加而非常迅速地增加。因此，其阶数越高，病态程度就越为严重。

定理1.2 设 $A\in\mathbb{R}^{n\times n}$ 非奇异，$b\in\mathbb{R}^n$ 非零。再假定 $\delta A\in\mathbb{R}^{n\times n}$ 满足 $\|A^{-1}\|\|\delta A\|<1$。若 x 和 $x+\delta x$ 分别是方程组

$$Ax=b \quad 和 \quad (A+\delta A)(x+\delta x)=b+\delta b$$

的解，则

$$\frac{\|\delta x\|}{\|x\|}\leqslant\frac{\kappa}{1-\kappa\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta A\|}{\|A\|}+\frac{\|\delta b\|}{\|b\|}\right),\tag{1.8}$$

其中 $\kappa=\kappa(A)=\|A\|\|A^{-1}\|$，矩阵范数是由相应的向量范数诱导出的算子范数。

证明 注意到

$$A+\delta A=(I+\delta AA^{-1})A$$

和

$$\|\delta AA^{-1}\|\leqslant\|\delta A\|\|A^{-1}\|<1,$$

据第一章的定理3.7和定理3.9知，$A+\delta A$ 可逆，而且

$$\|(I+\delta AA^{-1})\|\leqslant\frac{1}{1-\|\delta A\|\|A^{-1}\|}.\tag{1.9}$$

因此，

$$\begin{aligned}x+\delta x&=(A+\delta A)^{-1}(b+\delta b)\\&=(A+\delta A)^{-1}b+(A+\delta A)^{-1}\delta b.\end{aligned}$$

将 $x=A^{-1}b$ 代入上式，并移项，可得

$$\delta x=[(A+\delta A)^{-1}-A^{-1}]b+(A+\delta A)^{-1}\delta b.$$

上式两边取范数，可得

$$\|\delta x\|\leqslant\|[(A+\delta A)^{-1}-A^{-1}]b\|+\|(A+\delta A)^{-1}\|\|\delta b\|.\tag{1.10}$$

利用恒等式

$$(A+\delta A)^{-1}-A^{-1}=-A^{-1}(I+\delta AA^{-1})^{-1}\delta AA^{-1},$$

并注意到不等式（1.9），可得

$$\begin{aligned}\|[(A+\delta A)^{-1}-A^{-1}]b\|&=\|A^{-1}(I+\delta AA^{-1})^{-1}\delta Ax\|\\&\leqslant\|A^{-1}\|\|(I+\delta AA^{-1})^{-1}\|\|\delta A\|\|x\|\\&\leqslant\frac{\|\delta A\|\|A^{-1}\|}{1-\|\delta A\|\|A^{-1}\|}\|x\|.\end{aligned}\tag{1.11}$$

而利用不等式 (1.9)，又可得

$$\|(A+\delta A)^{-1}\| = \|A^{-1}(I+\delta A A^{-1})^{-1}\|$$

$$\leqslant \frac{\|A^{-1}\|}{1-\|\delta A\|\|A^{-1}\|}. \tag{1.12}$$

将(1.11)和(1.12)代入(1.10)，即得

$$\|\delta x\| \leqslant \frac{\|A^{-1}\|}{1-\|\delta A\|\|A^{-1}\|}[\|\delta A\|\|x\| + \|\delta b\|]. \tag{1.13}$$

(1.13) 两边同除以$\|x\|$，并注意到

$$\|b\| = \|Ax\| \leqslant \|A\|\|x\|,$$

就有

$$\frac{\|\delta x\|}{\|x\|} \leqslant \frac{\|A^{-1}\|\|A\|}{1-\|\delta A\|\|A^{-1}\|}\left[\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right],$$

即(1.8) 成立。

注1.1 定理1.2的证明过程亦证明了：如果 $A \in \mathbb{R}^{n\times n}$可逆，$\delta A \in \mathbb{R}^{n\times n}$满足$\|\delta A\|\|A^{-1}\|<1$，则 $A+\delta A$ 可逆，并且有

$$\frac{\|(A+\delta A)^{-1}-A^{-1}\|}{\|A\|} \leqslant \frac{\kappa\frac{\|\delta A\|}{\|A\|}}{1-\kappa\frac{\|\delta A\|}{\|A\|}}.$$

这表明$\kappa=\kappa(A)=\|A\|\|A^{-1}\|$亦可作为矩阵求逆问题的条件数。

§2 基本解法的回顾

设$A \in \mathbb{R}^{n\times n}$非奇异，$b \in \mathbb{R}^n$. 这一节我们来简要地复习一下求解线性方程组

$$Ax = b \tag{2.1}$$

的最基本解法。

2.1 Gauss 消去法

大家熟知，当 A 是稠密矩阵时，目前求解(2.1)的最有效方法是选主元素的 Gauss 消去法。因此，我们首先来复习这一方法的要点。

1. Gauss 消去法的基本步骤

(1) 利用 Gauss 变换求 A 的 LU 分解：$A = LU$，其中 L 是单位下三角阵，U 是上三角阵；

(2) 求解 $Ly = b$，得 y；

(3) 求解 $Ux = y$，得(2.1)的解 x。

这一方法虽然简单易行，然而它却有两个致命的弱点：

(1) 适用范围小。能够对 A 进行 LU 分解的前提是 A 从 1 到 $n-1$ 阶顺序主子式皆不为零。因此，像

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

这样一类非常良态（条件数 $\kappa_2(A) = 1$）的矩阵，也不存在 LU 分解。

(2) 数值稳定性差。误差分析的结果表明，按 Gauss 消去法计算得到的解 $\tilde{x}$ 满足

$$(A + E)\tilde{x} = b,$$

其中

$$|E| \leqslant n\varepsilon\{3|A| + 5|\hat{L}||\hat{U}|\} + O(\varepsilon^2),$$

这里 $\hat{L}$ 和 $\hat{U}$ 分别是 L 和 U 的计算结果，ε 是机器精度。由于其中的 $|\hat{L}||\hat{U}|$ 可以很大，因而数值稳定性差。

理论分析的结果表明，产生上述两个问题的主要原因是零主元素和小主元素的出现。因此，选主元素的 Gauss 消去法就随之而产生。

2. 全选主元素的 Gauss 消去法

（1）求排列方阵 P,Q 和分解：

$$PAQ = LU,$$

其中 U 是上三角矩阵，$L=[l_{ij}]$ 为满足 $|l_{ij}|\leqslant 1$ 的单位下三角矩阵。

（2）将(2.1)分解为四个简单易解的方程组进行求解。

这样做的结果是弥补了不选主元素的 Gauss 消去法的不足，然而付出的代价也是极其昂贵的。因为选主元素必须进行 $\sum_{k=1}^{n}k^2-n$ 次两个元素之间的比较和相应的逻辑判断，这在计算机上是相当费时间的。为了尽可能地减少所进行的比较，人们提出了所谓的部分选主元素的 Gauss 消去法。

3．部分选主元素的 Gauss 消去法

在全选主元素的 Gauss 消去法中，只需取 $Q=I$ 即可，即只在当前的列中选主元素，而不涉及其他元素。

实际计算的经验和理论分析的结果都表明，部分选主元素的 Gauss 消去法与全主元的 Gauss 消去法在数值稳定性方面完全可以媲美，但它的工作量却大为减少（只需进行 $(n-1)n/2$ 次两个元素之间的比较即可）。因此，它受到了人们的青睐，成为求解中小型稠密线性方程组最受欢迎的方法之一。

2.2 Cholesky 分解法

对于一般方阵，为了消除 LU 分解的局限性和误差的过分积累，而采用了选主元素的方法。但对于正定矩阵而言，选主元素却是完全不必要的。

设线性方程组(2.1)的系数矩阵 A 是对称正定的。此时，求解(2.1)的行之有效的方法是所谓的 Cholesky 分解法，其实质是不选主元素的 Gauss 消去法。Cholesky 分解法的基本步骤如下：

（1）求 A 的 Cholesky 分解：$A=GG^{\mathrm{T}}$，其中 G 是对角元素均为正数的下三角阵；

（2）求解 $Gy=b$，得 y；

（3）求解 $G^{T}x=y$，得 x。

在实际使用时，计算 A 的Cholesky 分解是采用如下方式进行的。

先验地记

$$G=\begin{bmatrix} g_{11} & & & 0 \\ g_{21} & g_{22} & & \\ \vdots & \vdots & \ddots & \\ g_{n1} & g_{n2} & \cdots & g_{nn} \end{bmatrix},$$

然后比较 $A=GG^{T}$ 两边对应的元素，得关系式

$$a_{ij}=\sum_{k=1}^{j} g_{ik}g_{jk}, \quad 1\leqslant j\leqslant i\leqslant n.$$

由这一公式，可以逐列求出 g_{ij}（从第一列开始）。

应用Cholesky 分解法求对称正定方程组，在数值稳定性方面完全可以与全选主元素的 Gauss 消去法媲美，而它的运算量却仅是 Gauss 消去法的一半，并且省去了元素之间的比较和相应的逻辑判断。因此，Cholesky 分解法是目前求解中小型稠密对称正定线性方程组的最佳方法之一。

§3 对称不定方程组的解法

上节已经讲过，用 Cholesky 分解法求对称正定方程组时，运算量仅是 Gauss 消去法的一半。那么，对于对称不定方程组，是否亦可利用对称性而给出与Cholesky 分解法运算量相同的数值解法呢？答案是肯定的。这一节，我们就来介绍这方面的一个较好的方法。

设 $A\in S\mathbb{R}^{n\times n}$ 非奇异，$b\in\mathbb{R}^{n}$。考虑线性方程组

$$Ax=b. \tag{3.1}$$

对于给定的 $A \in S\mathbb{R}^{n \times n}$，如果它有 LU 分解

$$A = LU, \tag{3.2}$$

其中 L 是单位下三角阵，U 是非奇异的上三角阵，则易知(3.2)可以写成如下形式

$$A = LDL^{\mathrm{T}}, \tag{3.3}$$

其中 D 是由 U 的对角元素构成的对角阵. 通常称(3.3)为对称矩阵的 $\mathrm{LDL^T}$ 分解. 这样，我们自然会想到通过求 A 的 $\mathrm{LDL^T}$ 分解来求解(3.1)。然而，在 A 非正定的情形下，这样做与不选主元素的 Gauss 消去法一样是不适用的，必须进行必要的主元素选取。但是，列选主元素或全选主元素势必要破坏 A 的对称性。因此，为保持对称性，我们必须对行列施行同样的对换，即应选取排列方阵 P，使

$$PAP^{\mathrm{T}} = LDL^{\mathrm{T}}, \tag{3.4}$$

其中 $L=[l_{ij}]$ 为满足 $|l_{ij}| \leqslant 1$ 的下三角阵，D 为对角阵. 不幸的是，一般来说这样的排列方阵并不一定存在。例如，

$$P\begin{bmatrix} 10^{-10} & 1 \\ 1 & 10^{-10} \end{bmatrix}P = \begin{bmatrix} 10^{-10} & 1 \\ 1 & 10^{-10} \end{bmatrix}$$

对一切的排列方阵都成立，但

$$\begin{bmatrix} 10^{-10} & 1 \\ 1 & 10^{-10} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 10^{10} & 1 \end{bmatrix}\begin{bmatrix} 10^{-10} & 0 \\ 0 & 10^{-10}-10^{10} \end{bmatrix}\begin{bmatrix} 1 & 10^{10} \\ 0 & 1 \end{bmatrix},$$

其中的单位下三角阵的左下角元素 $10^{10} \gg 1$。

基于上述原因，Aasen (1971)提出了求解(3.1)的如下方案：

(1) 求分解

$$PAP^{\mathrm{T}} = LTL^{\mathrm{T}}, \tag{3.5}$$

其中 $L=[l_{ij}]$ 为满足 $|l_{ij}| \leqslant 1$ 的单位下三角阵，P 是排列方阵，

T 是对称三对角阵；

(2) 解方程组 $Lw=Pb$, $Tz=w$, $L^{\mathrm{T}}v=z$ 和 $Px=v$。

实现这一方案的关键在于实现分解(3.5)。至于(2)中的四个方程组，除 $Tz=w$ 外，都是很容易求解的。而对于 $Tz=w$,由于 T 是对称三对角矩阵,故可不管其对称性利用部分选主元的Gauss消去法仅用 $O(n)$次运算就可求得它的解。因此，下面我们就致力于分解式(3.5)的实现。

类比于LU分解的实现过程,我们自然想到利用稳定的Gauss变换将 A 逐步约化为对称三对角阵而实现分解式(3.5)。

记 $T_0=A$，并假定对某个自然数 k，$1\leqslant k\leqslant n-2$，我们已求得 $k-1$ 个排列方阵 $P_1,\cdots,P_{k-1}$ 和 $k-1$ 个 Gauss 变换 $M_1,\cdots,M_{k-1}$，使得

$$M_{k-1}P_{k-1}\cdots M_1P_1AP_1^{\mathrm{T}}M_1^{\mathrm{T}}\cdots P_{k-1}^{\mathrm{T}}M_{k-1}^{\mathrm{T}}=T_{k-1}, \tag{3.6}$$

其中 T_{k-1} 具有如下形状

$$T_{k-1}=\begin{bmatrix} T_{11}^{(k-1)} & T_{12}^{(k-1)} \\ T_{21}^{(k-1)} & T_{22}^{(k-1)} \end{bmatrix}\begin{matrix} k \\ n-k \end{matrix}, \tag{3.7}$$
$$\qquad\qquad k \qquad\quad n-k$$

这里 $T_{22}^{(k-1)}\in S\mathbb{R}^{(n-k)\times(n-k)}$，$T_{11}^{(k-1)}$ 是如下的对称三对角矩阵：

$$T_{11}^{(k-1)}=\begin{bmatrix} \alpha_1 & \beta_2 & & 0 \\ \beta_2 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_k \\ 0 & & \beta_k & \alpha_k \end{bmatrix}, \tag{3.8}$$

$$(T_{12}^{(k-1)})^{\mathrm{T}}=T_{21}^{(k-1)}=[\underset{k-1}{0},\underset{1}{v_{k-1}}]\in\mathbb{R}^{(n-k)\times k}, \tag{3.9}$$

$$v_{k-1}=(\zeta_1^{(k-1)},\cdots,\zeta_{n-k}^{(k-1)})^{\mathrm{T}}\in\mathbb{R}^{n-k}.$$

我们的第 k 步是：

(i) 确定指标 q，使得

$$|\zeta_q^{(k-1)}|=\max\{|\zeta_1^{(k-1)}|,\cdots,|\zeta_{n-k}^{(k-1)}|\}; \tag{3.10}$$

(ii) 取 $\tilde{P}_k$ 为第一行与第 q 行交换的 $n-k$ 阶初等交换矩阵，

并记

$$\tilde{v}_{k-1}=(\tilde{\xi}_1^{(k-1)},\cdots,\tilde{\xi}_{n-k}^{(k-1)})^{\mathrm{T}}\doteq\tilde{P}_k v_{k-1},\tag{3.11}$$

$$P_k=\operatorname{diag}(I_k,\tilde{P}_k);$$

(iii) 取 $\widetilde{M}_k$ 为 $n-k$ 阶 Gauss 变换

$$\widetilde{M}_k=I_{n-k}-(0,\tilde{l}_{k+2,k+1}^{(k)},\cdots,\tilde{l}_{n,k+1}^{(k)})^{\mathrm{T}}[e_1^{(n-k)}]^{\mathrm{T}}$$

$$=\begin{bmatrix}1 & & & 0\\ -\tilde{l}_{k+2,k+1}^{(k)} & 1 & & \\ -\tilde{l}_{k+3,k+1}^{(k)} & 0 & \ddots & \\ \vdots & \vdots & \ddots & \ddots \\ -\tilde{l}_{n,k+1}^{(k)} & 0 & \cdots & 0 \quad 1\end{bmatrix},\tag{3.12}$$

其中 $\tilde{l}_{k+i,k+1}^{(k)}=\tilde{\xi}_i^{(k-1)}/\tilde{\xi}_1^{(k-1)}$，$i=2,3,\cdots,n-k$，并令

$$M_k=\operatorname{diag}(I_k,\widetilde{M}_k);$$

(iv) 计算

$$\begin{matrix}1\\ n-k-1\end{matrix}\begin{bmatrix}\alpha_{k+1} & v_k^{\mathrm{T}}\\ v_k & T_{22}^{(k)}\end{bmatrix}=\widetilde{M}_k\tilde{P}_k T_{22}^{(k-1)}\tilde{P}_k^{\mathrm{T}}\widetilde{M}_k^{\mathrm{T}}.\tag{3.13}$$

$$\qquad\quad 1 \qquad n-k-1$$

如果我们记

$$T_k=\begin{bmatrix}T_{11}^{(k)} & T_{12}^{(k)}\\ T_{21}^{(k)} & T_{22}^{(k)}\end{bmatrix}\begin{matrix}k+1\\ n-k-1\end{matrix},\tag{3.14}$$

$$\qquad k+1 \quad n-k-1$$

其中

$$T_{11}^{(k)}=\begin{bmatrix}\alpha_1 & \beta_2 & & 0\\ \beta_2 & \ddots & \ddots & \\ & \ddots & \alpha_k & \beta_{k+1}\\ 0 & & \beta_{k+1} & \alpha_{k+1}\end{bmatrix},\qquad \beta_{k+1}=\tilde{\xi}_1^{(k-1)},$$

$$(T_{12}^{(k)})^{\mathrm{T}}=T_{21}^{(k)}=[\underset{k}{0},\underset{1}{v_k}]\in\mathbb{R}^{(n-k-1)\times(k+1)},$$

则第 k 步就相当于把 T_{k-1} 约化为 T_k，即

$$M_k P_k T_{k-1} P_k^{\mathrm{T}} M_k^{\mathrm{T}} = T_k. \tag{3.15}$$

这样，从 T_0 出发，经过 $n-2$ 步就可把 A 约化为对称三对角矩阵。现在令

$$L_1 = M_1^{-1}, \quad A_0 = A, \tag{3.16}$$

$$L_k = P_k L_{k-1} P_k M_k^{-1}, \quad k = 2, \cdots, n-2, \tag{3.17}$$

$$A_k = P_k A_{k-1} P_k, \quad k = 1, 2, \cdots, n-2. \tag{3.18}$$

$$L = L_{n-2}, \quad T = T_{n-2}, \quad P = P_{n-2} \cdots P_1. \tag{3.19}$$

则易知(3.15)用现在的记号即可写成

$$A_k = L_k T_k L_k^{\mathrm{T}}, \quad k = 1, 2, \cdots, n-2; \tag{3.20}$$

特别，对 $k = n-2$，有

$$PAP^{\mathrm{T}} = A_{n-2} = L_{n-2} T_{n-2} L_{n-2}^{\mathrm{T}} = LTL^{\mathrm{T}}.$$

因此，只要证明了 $L = L_{n-2}$ 是所有元素的绝对值均小于 1 的单位下三角阵，则我们就已求得了分解式(3.5)。事实上，从 L_{k-1} 的定义，用归纳法可证它具有如下形状

$$L_{k-1} = \underset{\quad k \qquad\quad n-k}{\begin{bmatrix} L_{11}^{(k-1)} & 0 \\ L_{21}^{(k-1)} & I_{n-k} \end{bmatrix}} \begin{matrix} k \\ n-k \end{matrix} \quad (k = 2, \cdots, n-1), \tag{3.21}$$

其中

$$L_{11}^{(k-1)} = \begin{bmatrix} 1 & & & & \\ 0 & 1 & & \text{0} & \\ 0 & l_{32}^{(k-1)} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ 0 & l_{k2}^{(k-1)} & \cdots & l_{k,k-1}^{(k-1)} & 1 \end{bmatrix},$$

$$L_{21}^{(k-1)} = \begin{bmatrix} 0 & l_{k+1,2}^{(k-1)} & \cdots & l_{k+1,k}^{(k-1)} \\ 0 & l_{k+2,2}^{(k-1)} & \cdots & l_{k+2,k}^{(k-1)} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & l_{n,2}^{(k-1)} & \cdots & l_{nk}^{(k-1)} \end{bmatrix},$$

且 $|l_{ij}^{(k-1)}|\leqslant 1$。

由 $L_1=M_1^{-1}$ 和 M_1 的定义立即知(3.21)对 $k=2$ 成立。现假定对 $k-1$ 已证(3.21)成立，则

$$
\begin{aligned}
L_k &= P_k L_{k-1} P_k M_k^{-1} \\
&= \begin{bmatrix} I_k & 0 \\ 0 & \tilde{P}_k \end{bmatrix} \begin{bmatrix} L_{11}^{(k-1)} & 0 \\ L_{21}^{(k-1)} & I_{n-k} \end{bmatrix} \begin{bmatrix} I_k & 0 \\ 0 & \tilde{P}_k \end{bmatrix} \begin{bmatrix} I_k & 0 \\ 0 & \tilde{M}_k^{-1} \end{bmatrix} \\
&= \begin{bmatrix} L_{11}^{(k-1)} & 0 \\ \tilde{P}_k L_{21}^{(k-1)} & \tilde{M}_k^{-1} \end{bmatrix}.
\end{aligned}
\tag{3.22}
$$

而

$$
\tilde{M}_k^{-1} = \begin{bmatrix} 1 & & & & 0 \\ \tilde{l}_{k+2,k+1}^{(k)} & 1 & & & \\ \vdots & 0 & \ddots & & \\ \vdots & \vdots & \ddots & \ddots & \\ \tilde{l}_{n,k+1}^{(k)} & 0 & \cdots & 0 & 1 \end{bmatrix},
$$

且 $|\tilde{l}_{k+i,k+1}^{(k)}|=|\xi_i^{(k-1)}/\xi_1^{(k-1)}|\leqslant 1$，故(3.22)表明(3.21)对 k 亦成立。于是，由归纳法原理知，(3.21)对 $k=2,\cdots,n-2$ 都成立。

顺便指出，(3.22)亦说明我们可从 $L_1=M_1^{-1}$ 出发，一步步很容易地把 $L=L_{n-2}$ 求出来。

现在，我们来看这样实现分解式(3.5)的运算量是多少。从上述实现过程可以看出，每步的主要工作量是(3.13)的计算。而完成这一计算最经济的方法也需运算量$(n-k)^2$。因此，整个过程所需的运算量是 $n^3/3$，这与 Gauss 消去法的运算量一样，而比我们所希望的运算量多出一倍。

仔细观察上述过程可知，其第 k 步确定排列方阵 P_k 和Gauss 变换 M_k 时，只涉及到向量 v_{k-1} 的信息，而并不需要 $T_{22}^{(k-1)}$ 的信息。Aasen 正是注意到这一点，给出了不需明确求出 $T_{22}^{(k-1)}$，而直接计算 v_{k-1} 的方法，从而使运算量大为减少。

现在假定 L_{k-1} 和 A_{k-1} 已确定，并假定 T_{k-1} 左上角的对称

三对角阵 $T_{11}^{(k-1)}$ 除 α_k 之外其他元素都已确定。我们来考虑如何根据这些已知信息来求出 v_{k-1} 和 α_k 及 β_{k+1}。

从(3.20)可得

$$
\begin{aligned}
A_{k-1} &= L_{k-1}T_{k-1}L_{k-1}^{\mathrm{T}} \\
&= \begin{bmatrix} L_{11}^{(k-1)} & 0 \\ L_{21}^{(k-1)} & I_{n-k} \end{bmatrix}\begin{bmatrix} T_{11}^{(k-1)} & T_{12}^{(k-1)} \\ T_{21}^{(k-1)} & T_{22}^{(k-1)} \end{bmatrix}\begin{bmatrix} (L_{11}^{(k-1)})^{\mathrm{T}} & (L_{21}^{(k-1)})^{\mathrm{T}} \\ 0 & I_{n-k} \end{bmatrix} \\
&= \begin{bmatrix} L_{11}^{(k-1)}T_{11}^{(k-1)}(L_{11}^{(k-1)})^{\mathrm{T}} & * \\ (L_{21}^{(k-1)}T_{11}^{(k-1)} + T_{21}^{(k-1)})(L_{11}^{(k-1)})^{\mathrm{T}} & * \end{bmatrix}.
\end{aligned}
\tag{3.23}
$$

注意到 $T_{21}^{(k-1)} = [0, v_{k-1}]$，而 $L_{11}^{(k-1)}$ 是单位下三角阵，即知

$$T_{21}^{(k-1)}(L_{11}^{(k-1)})^{\mathrm{T}} = T_{21}^{(k-1)}.$$

这样，比较(3.23)两边矩阵的第 k 列的后 $n-k$ 个元素，可得

$$\begin{bmatrix} a_{k+1,k}^{(k-1)} \\ a_{k+2,k}^{(k-1)} \\ \vdots \\ a_{nk}^{(k-1)} \end{bmatrix} = L_{21}^{(k-1)}T_{11}^{(k-1)}(L_{11}^{(k-1)})^{\mathrm{T}}e_k^{(k)} + v_{k-1}. \tag{3.24}$$

记

$$T_{11}^{(k-1)}(L_{11}^{(k-1)})^{\mathrm{T}}e_k^{(k)} = (\xi_1, \cdots, \xi_k)^{\mathrm{T}}, \tag{3.25}$$

则(3.24)可写成

$$\zeta_i^{(k-1)} = a_{k+i,k}^{(k-1)} - \sum_{j=2}^{k} l_{k+i,j}^{(k-1)}\xi_j, \quad i = 1, 2, \cdots, n-k. \tag{3.26}$$

再比较(3.25)两边的元素，可得

$$\begin{cases} \xi_1 = \beta_2 l_{k2}^{(k-1)}, \\ \xi_i = \beta_i l_{k,i-1}^{(k-1)} + \alpha_i l_{k,i}^{(k-1)} + \beta_{i+1} l_{k,i+1}^{(k-1)}, \quad i = 2, \cdots, k-1, \\ \xi_k = \beta_k l_{k,k-1}^{(k-1)} + \alpha_k. \end{cases} \tag{3.27}$$

注意公式(3.27)中含有待求的量α_k。因此，由(3.27)还不能确定ξ_k。幸运的是，比较(3.23)两边矩阵的第 k 个对角元素，可得

$$\xi_k = a_{kk}^{(k-1)} - \sum_{j=2}^{k-1} l_{kj}^{(k-1)}\xi_j. \tag{3.28}$$

这样，我们就可根据已知信息，由(3.27)的前 $k-1$ 个等式确定 $\xi_1,\cdots,\xi_{k-1}$；再由(3.28)确定 ξ_k；然后再由(3.26)就可确定向量 v_{k-1} 的 $n-k$ 个分量。

另外，(3.27)的最后一个等式，也给我们提供了计算 α_k 的公式

$$\alpha_k=\begin{cases}\alpha_{11}, & k=1,\\ \xi_k-\beta_k l_{k,k-1}^{(k-1)}, & k>1,\end{cases} \tag{3.29}$$

其中 α_{11} 是 A 的第一个对角元素。

求出 v_{k-1} 之后，我们就可由(3.10)，(3.11)和(3.12)确定 v_{k-1} 的最大分量 $\zeta_q^{(k-1)}$，排列方阵 $\tilde{P}_k$ 和 Gauss 变换 $\tilde{M}_k$，同时也确定了 $\beta_{k+1}=\zeta_q^{(k-1)}$。然后由(3.18)和(3.22)求得 A_k 和 L_k。如此进行 $n-2$ 步之后，就可从 $A_0=A$ 出发递推地求得 T，L 和 P，而且容易算出这样实现分解(3.5)的全部工作量是 $n^3/6$。

此外，在实际计算时，我们可将 L 存放在 A 的下三角部分，P 以因子形式存放（即存放 P_i 的对换指标），T 的对角元素和次对角元素分别存放在 A 的相应位置上。例如，$n=6$时，计算得到的 L 和 T 有如下形状

$$L=\begin{bmatrix}1 & & & & & \\ 0 & 1 & & & & \\ 0 & l_{32} & 1 & & & \\ 0 & l_{42} & l_{43} & 1 & & \\ 0 & l_{52} & l_{53} & l_{54} & 1 & \\ 0 & l_{62} & l_{63} & l_{64} & l_{65} & 1\end{bmatrix},$$

$$T=\begin{bmatrix}\alpha_1 & \beta_2 & & & & \\ \beta_2 & \alpha_2 & \beta_3 & & & \\ & \beta_3 & \alpha_3 & \beta_4 & & \\ & & \beta_4 & \alpha_4 & \beta_5 & \\ & & & \beta_5 & \alpha_5 & \beta_6 \\ & & & & \beta_6 & \alpha_6\end{bmatrix}.$$

我们可按如下方式存放 L 和 T：

$$\begin{bmatrix} \alpha_1 & & & & & \\ \beta_2 & \alpha_2 & & & & \\ l_{32} & \beta_3 & \alpha_3 & & & \\ l_{42} & l_{43} & \beta_4 & \alpha_4 & & \\ l_{52} & l_{53} & l_{54} & \beta_5 & \alpha_5 & \\ l_{62} & l_{63} & l_{64} & l_{65} & \beta_6 & \alpha_6 \end{bmatrix}.$$

由上面的讨论和所设计的存储方案，可得求分解式(3.5)的算法如下：

算法3.1

(1) 输入 $A=[a_{ij}]$； $k:=1$。

(2) $l_1:=0$,

$l_i:=a_{k,i-1} \quad (i=2,\cdots,k-1)$,

$l_k:=1$,

$\xi_i:=a_{i,i-1}l_{i-1}+a_{ii}l_i+a_{i+1,i}l_{i+1} \quad (i=2,\cdots,k-1)$,

$\xi_k:=a_{kk}-\sum_{k=2}^{k-1}l_i\xi_i$.

(3) 如果 $k>1$，则 $a_{kk}:=\xi_k-a_{k,k-1}l_{k-1}$.

(4) 如果 $k<n$，则

$$\zeta_i:=a_{ik}-\sum_{j=2}^{k}a_{i,j-1}\xi_j \quad (i=k+1,\cdots,n);$$

否则转步(9)。

(5) 如果 $k<n-1$，则确定下标 $q(k+1\leqslant q\leqslant n)$，使得

$$|\zeta_q|=\max\{|\zeta_{k+1}|,\cdots,|\zeta_n|\},$$

且 $p_k:=q$（记录交换阵 P_k）；否则转步(8)。

(6) 交换 ζ_{k+1} 和 ζ_q， $a_{k+1,j}$ 和 $a_{qj}(j=1,\cdots,n)$，以及 $a_{j,k+1}$ 和 a_{jq} $(j=k+1,\cdots,n)$。

(7) 如果 $\zeta_{k+1}\neq 0$，则

$$\alpha_{ik}:=\zeta_i/\zeta_{k+1} \quad (i=k+1,\cdots,n);$$

否则

$$\alpha_{ik}:=0 \quad (i=k+2,\cdots,n).$$

(8) $\alpha_{k+1,k}:=\zeta_{k+1}$，$k:=k+1$，转步(2)。

(9) 输出有关信息，结束。

这一算法本质上与部分选主元素的 LU 分解是一样的，因此它是数值稳定的，但它的运算量却仅是 LU 分解的一半。

§4 Vandermonde方程组的解法

这一节，我们来考虑 Vandermonde 方程组

$$Vz=b \tag{4.1}$$

的求解问题，其中 $z=(z_0,z_1,\cdots,z_n)^T\in\mathbb{R}^{n+1}$ 是未知向量，$b=(\beta_0,\cdots,\beta_n)^T\in\mathbb{R}^{n+1}$ 是已知向量，V 是Vandermonde 矩阵

$$V=V(x_0,\cdots,x_n)=\begin{bmatrix}1 & 1 & \cdots & 1\\ x_0 & x_1 & \cdots & x_n\\ x_0^2 & x_1^2 & \cdots & x_n^2\\ \vdots & \vdots & & \vdots\\ x_0^n & x_1^n & \cdots & x_n^n\end{bmatrix}, \tag{4.2}$$

这里 $x_0,x_1,\cdots,x_n$ 是给定的 $n+1$ 个互不相同的实数。

注意到Vandermonde 方程组与多项式插值之间的密切关系，我们就可设计出运算量仅为 n^2 的求解(4.1)的数值方法。

大家知道，多项式插值就是，给定函数 $y=f(x)$在 $n+1$ 个不同点 x_i 处的函数值 y_i $(i=0,1,2,\cdots,n)$，要求构造一个次数不超过 n 的多项式

$$p(x)=a_nx^n+a_{n-1}x^{n-1}+\cdots+a_1x+a_0,$$

使得 $p(x)$在节点 x_i 处满足

$$p(x_i)=f(x_i)=y_i, \quad i=0,1,\cdots,n. \tag{4.3}$$

将其用矩阵向量的语言表述出来就是，求向量 $v=(a_0,\cdots,a_n)^{\mathrm{T}}\in \mathrm{R}^{n+1}$，使得

$$V^{\mathrm{T}}v=y, \tag{4.4}$$

其中 V 是形如(4.2)的 Vandermonde 矩阵，$y=(y_0,\cdots,y_n)^{\mathrm{T}}$。

根据插值多项式的理论知，存在唯一的多项式满足（4.3)，并且 $p(x)$的系数 $a_0,a_1,\cdots,a_n$ 可分两步求得：

（1）求 $p(x)$的 Newton 表达式

$$p(x)=c_0+\sum_{k=1}^{n}c_k\prod_{i=0}^{k-1}(x-x_i); \tag{4.5}$$

（2）将(4.5)按 x 的幂展开，即得 $p(x)$的系数 $a_0,a_1,a_2,\cdots,a_n$。

根据 Newton 插值多项式的理论知，(4.5)的系数 c_k 就是$p(x)$在 $x_0,x_1,\cdots,x_k$ 处的 k 阶**差商**。差商又称**均差**，其定义如下：

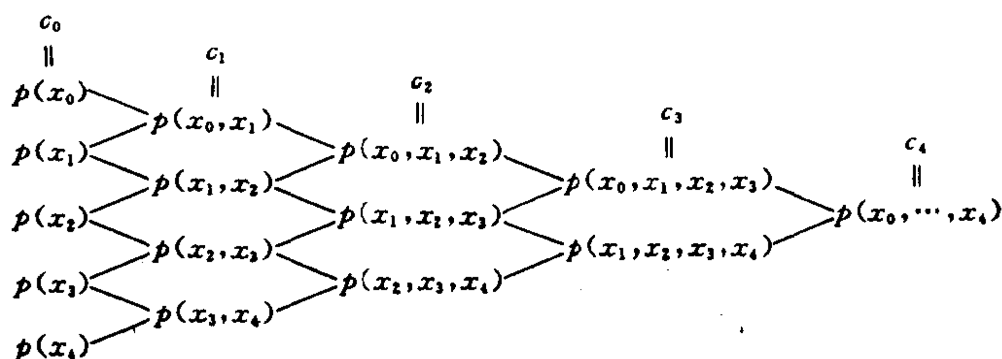
$p(x)$在 x_i,x_j 处的一阶差商定义为

$$p(x_i,x_j)=\frac{p(x_i)-p(x_j)}{x_i-x_j},\quad x_i\neq x_j;$$

$p(x)$在 $x_0,\cdots,x_l$ 这 $l+1$ 个互异点处的 l 阶差商定义为

$$p(x_0,\cdots,x_l)=\frac{p(x_0,\cdots,x_{l-1})-p(x_1,x_2,\cdots,x_l)}{x_0-x_l}.$$

因此，c_k 可以递推地求得。例如，当$n=4$ 时，计算过程可列表如下：



一般地，记

$$u_k=(\xi_0^{(k)},\cdots,\xi_n^{(k)})^{\mathrm{T}},\quad k=0,1,\cdots,n,$$

其中

$$\xi_i^{(k)}=\begin{cases} c_i, & 0\leqslant i\leqslant k, \\ p(x_{i-k},\cdots,x_i), & k+1\leqslant i\leqslant n, \end{cases}$$

则计算过程的第 k 步就是将向量 u_{k-1} 变换为 u_k，即

$$u_k=D_k^{-1}L_k u_{k-1},\quad k=1,2,\cdots,n,$$

其中

$$D_k=\mathrm{diag}(1,\cdots,1,x_k-x_0,x_{k+1}-x_1,\cdots,x_n-x_{n-k}),$$

$$L_k=\begin{bmatrix} I_{k-1} & 0 \\ 0 & \tilde{L}_k \end{bmatrix}\in\mathbb{R}^{(n+1)\times(n+1)},$$

$$\tilde{L}_k=\begin{bmatrix} 1 & & & 0 \\ -1 & 1 & & \\ & \ddots & \ddots & \\ 0 & & -1 & 1 \end{bmatrix}\in\mathbb{R}^{(n-k+2)\times(n-k+2)}.$$

于是

$$\begin{aligned} c=u_n&=D_n^{-1}L_n D_{n-1}^{-1}L_{n-1}\cdots D_1^{-1}L_1 u_0 \\ &=D_n^{-1}L_n\cdots D_1^{-1}L_1 y, \end{aligned}\tag{4.6}$$

其中 $c=(c_0,c_1,\cdots,c_n)^{\mathrm{T}}$，$y=(p(x_0),\cdots,p(x_n))^{\mathrm{T}}=(y_0,\cdots,y_n)^{\mathrm{T}}$.

求出 $p(x)$的 Newton 表达式(4.5)之后，可按如下方式递推地将$p(x)$展成 x 的幂的形式：

$$\begin{aligned} &p_n(x)\equiv c_n, \\ &p_k(x)=c_k+(x-x_k)p_{k+1}(x),\quad k=n-1,\cdots,1,0, \end{aligned}\tag{4.7}$$

即由 c_n 出发，按(4.7)递推地计算 n 次，就可求得 $p(x)=p_0(x)$ 的按 x 的幂展开的形式，从而也就得到了其系数 $a_0,a_1,\cdots,a_n$.

记

$$p_k(x)=a_k^{(k)}+a_{k+1}^{(k)}x+\cdots+a_n^{(k)}x^{n-k},$$

比较(4.7)两边的x的同次幂的系数，可得计算$a_i^{(k)}$的递推公式：

$$\begin{cases} a_n^{(n)} = c_n, \\ a_k^{(k)} = c_k - x_k a_{k+1}^{(k+1)}, \\ a_i^{(k)} = a_i^{(k+1)} - x_k a_{i+1}^{(k+1)}, \quad i = k+1, \cdots, n-1, \\ a_n^{(k)} = a_n^{(k+1)}, \end{cases} \tag{4.8}$$

$$(k = n-1, n-2, \cdots, 1, 0.)$$

令

$$v_k = (\zeta_0^{(k)}, \cdots, \zeta_n^{(k)})^{\mathrm{T}}, \quad k = 0, 1, \cdots, n,$$

其中

$$\zeta_i^{(k)} = \begin{cases} c_i, & 0 \leqslant i \leqslant k-1, \\ a_i^{(k)}, & k \leqslant i \leqslant n. \end{cases}$$

再令

$$U_k = \left[\begin{array}{c|cccc} I_k & & & 0 & \\ \hline & 1 & -x_k & & 0 \\ & & \ddots & \ddots & \\ 0 & 0 & & \ddots & -x_k \\ & & & & 1 \end{array}\right] \begin{array}{l} k \\ \\ n-k+1 \end{array}, \quad k = n-1, \cdots, 0.$$

$$\begin{array}{cc} \quad k & \qquad n-k+1 \end{array}$$

则迭代公式(4.8)可写成

$$v_k = U_k v_{k+1}, \quad k = n-1, n-2, \cdots, 1, 0.$$

这样，$p(x) = p_0(x)$的系数作成的向量$v = (a_0, \cdots, a_n)^{\mathrm{T}}$就可表成

$$v = U_0 U_1 \cdots U_{n-1} v_n = U_0 U_1 \cdots U_{n-1} c. \tag{4.9}$$

因此，由(4.6)和(4.9)就可得到

$$v = ULy,$$

其中

$$U = U_0 U_1 \cdots U_{n-1}, \quad L = D_n^{-1} L_n \cdots D_1^{-1} L_1.$$

另一方面，从(4.4)可知，$v = V^{-\mathrm{T}} y$. 于是便有

$$V^{-T}y = ULy。\tag{4.10}$$

注意到这里的 $y=(y_0,\cdots,y_n)$是可以任意 指定的，即知(4.10)蕴含着

$$V^{-T} = UL,$$

即

$$V^{-1} = L^T U^T。$$

这表明我们已经求得了 Vandermonde 矩阵 V 的逆矩阵 V^{-1}的因式分解 $L^T U^T$。由此我们就可很容易地导出(4.1)的解为

$$\begin{aligned} z &= V^{-1}b = L^T U^T b \\ &= [D_n^{-1}L_n\cdots D_1^{-1}L_1]^T[U_0\cdots U_{n-1}]^T b \\ &= [L_1^T D_1^{-1}\cdots L_n^T D_n^{-1}][U_{n-1}^T\cdots U_0^T b]。\end{aligned}$$

再根据 L_k, D_k 和 U_k 的定义，立即可得求解(4.1)的如下 算法。

算法4.1

(1) 输入原始数据：$\beta_0,\cdots,\beta_n$； $x_0,\cdots,x_n$； $k:=0$。

(2) $\beta_i:=\beta_i - x_k\ \beta_{i-1}\quad(i=n,\cdots,k+1)$。

(3) 如果 $k<n-1$，则 $k:=k+1$，转 步(2)；否则 进行下一步。

(4) $\beta_i:=\beta_i/(x_i - x_{i-k-1})\quad(i=k+1,\cdots,n)$；

$$\beta_i:=\beta_i-\beta_{i+1}\quad(i=k,\cdots,n-1)。$$

(5) 如果 $k>0$，则 $k:=k-1$，转 步(4)；否则，输 出有关信息，结束。

这一算法的运算量为 n^2，且将方程组的解存放在 β_i 的 存储单元里。

Björck–Pereyra（参见文献[19]）曾对这一算 法作过详 细的分析。他们的结果表明：这一算法具有较好的数值性态。

§5 Toeplitz方程组的解法

设 $A=[a_{ij}]\in\mathbb{R}^{n\times n}$。如果存在常数 $\gamma_{1-n},\gamma_{2-n},\cdots,\gamma_{-1},\gamma_0,\gamma_1,\cdots,\gamma_{n-1}$，使得

$$a_{ij}=\gamma_{j-i},\qquad i,j=1,2,\cdots,n,$$

则称 A 是**Toeplitz 矩阵**；即如果 A 是 Toeplitz 矩阵，则它具有如下形状

$$A=\begin{bmatrix}\gamma_0 & \gamma_1 & \cdots & \cdots & \gamma_{n-1}\\ \gamma_{-1} & \gamma_0 & \ddots & & \vdots\\ \vdots & \ddots & \ddots & \ddots & \vdots\\ \vdots & & \ddots & \ddots & \gamma_1\\ \gamma_{1-n} & \cdots & \cdots & \gamma_{-1} & \gamma_0\end{bmatrix}.$$

由此可见，Toeplitz 矩阵关于它的东北-西南对角线是对称的。具有这样对称性的矩阵通常称作**广对称矩阵**，即若 $B=[\beta_{ij}]\in\mathbb{R}^{n\times n}$ 是广对称的，则它满足

$$\beta_{ij}=\beta_{n-j+1,n-i+1},\qquad i,j=1,2,\cdots,n;$$

这等价于 B 满足

$$B=EB^{\mathrm{T}}E,$$

其中

$$E=[e_n,e_{n-1},\cdots,e_1]$$

是 n 阶**反序单位矩阵**。由广对称矩阵的等价定义，易证：非奇异的广对称矩阵的逆亦是广对称的。

在这一节里，我们假定 $T_n\in\mathbb{R}^{n\times n}$ 是给定的对称正定的 Toeplitz 矩阵。不失一般性，可假定 T_n 具有如下形状

$$T_n=\begin{bmatrix}1 & \gamma_1 & \gamma_2 & \cdots & \gamma_{n-1}\\ & 1 & \gamma_1 & \ddots & \vdots\\ & & \ddots & \ddots & \gamma_2\\ & \text{对称} & & \ddots & \gamma_1\\ & & & & 1\end{bmatrix}. \qquad (5.1)$$

而且在下面的讨论中，我们将用 T_k 表示 T_n 的 k 阶顺序主子阵，即

$$T_k = \begin{bmatrix} 1 & \gamma_1 & \cdots & \gamma_{k-1} \\ \gamma_1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \gamma_1 \\ \gamma_{k-1} & \cdots & \gamma_1 & 1 \end{bmatrix} \in \mathbb{R}^{k \times k}, \qquad k = 1, 2, \cdots, n-1.$$

显然，T_k 亦是对称正定的 Toeplitz 矩阵。

下面我们分三部分来讨论系数矩阵为 T_n 的线性方程组的求解问题和 T_n^{-1} 的计算问题。

5.1 Yule-Walker 方程组

我们先来考虑一类特殊的 Toeplitz 方程组

$$T_n y = -(\gamma_1, \cdots, \gamma_{n-1}, \gamma_n)^{\mathrm{T}}, \tag{5.2}$$

其中的 $\gamma_1, \cdots, \gamma_{n-1}$ 就是(5.1)中确定矩阵 T_n 的 $n-1$ 个常数，γ_n 是任意给定的实数。这类方程组称作 **Yule-Walker 方程组**。

由于这类方程组之右端项的特殊性，所以若记 y_k 为 k 阶 Yule-Walker 方程组

$$T_k y_k = -(\gamma_1, \cdots, \gamma_k)^{\mathrm{T}} \quad (k = 1, 2, \cdots, n) \tag{5.3}$$

之解，则可导出由 y_k 确定 y_{k+1} 的关系式。为此，记

$$y_{k+1} = \begin{bmatrix} z_k \\ \alpha_k \end{bmatrix} \begin{matrix} k \\ 1 \end{matrix}, \quad r_k = (\gamma_1, \cdots, \gamma_k)^{\mathrm{T}}.$$

则 $T_{k+1} y_{k+1} = -(\gamma_1, \cdots, \gamma_k, \gamma_{k+1})^{\mathrm{T}}$ 可写作

$$\begin{bmatrix} T_k & E_k r_k \\ r_k^{\mathrm{T}} E_k & 1 \end{bmatrix} \begin{bmatrix} z_k \\ \alpha_k \end{bmatrix} = -\begin{bmatrix} r_k \\ \gamma_{k+1} \end{bmatrix},$$

即有

$$T_k z_k + \alpha_k E_k r_k = -r_k, \tag{5.4}$$

$$r_k^{\mathrm{T}} E_k z_k + \alpha_k = -\gamma_{k+1}, \tag{5.5}$$

其中 E_k 表示 k 阶反序单位矩阵。

注意到 T_k 是对称正定的 Toeplitz 矩阵蕴含着 $T_k^{-1} E_k = E_k T_k^{-1}$，从 (5.4)就可得

$$\begin{aligned} z_k &= T_k^{-1}(-r_k - \alpha_k E_k r_k) \\ &= y_k + \alpha_k E_k y_k. \end{aligned} \tag{5.6}$$

将(5.6)代入(5.5)，并移项整理，得

$$(1 + r_k^{\mathrm{T}} y_k)\alpha_k = -\gamma_{k+1} - r_k^{\mathrm{T}} E_k y_k. \tag{5.7}$$

注意到

$$\begin{bmatrix} I_k & E_k y_k \\ 0 & 1 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} T_k & E_k r_k \\ r_k^{\mathrm{T}} E_k & 1 \end{bmatrix} \begin{bmatrix} I_k & E_k y_k \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} T_k & 0 \\ 0 & 1 + r_k^{\mathrm{T}} y_k \end{bmatrix}$$

和 T_{k+1} 的正定性，即知 $1 + r_k^{\mathrm{T}} y_k > 0$。因此，在(5.7)两边同除以 $1 + r_k^{\mathrm{T}} y_k$，即得

$$\alpha_k = -(\gamma_{k+1} + r_k^{\mathrm{T}} E_k y_k)/(1 + r_k^{\mathrm{T}} y_k). \tag{5.8}$$

这样，我们就找到了 y_k 与 y_{k+1} 之间的关系。从而，我们就可从一阶 Yule-Walker 方程组的解 y_1 出发，利用公式(5.8)和(5.6)递推地求得方程组(5.2)之解；而且容易算出这一求解过程所需的运算量为$3n^2/2$。

此外，令

$$\delta_k = 1 + r_k^{\mathrm{T}} y_k, \qquad k = 1, 2, \cdots, n-1, \tag{5.9}$$

则有

$$\begin{aligned} \delta_{k+1} &= 1 + r_{k+1}^{\mathrm{T}} y_{k+1} \\ &= 1 + (r_k^{\mathrm{T}}, \gamma_{k+1}) \begin{bmatrix} y_k + \alpha_k E_k y_k \\ \alpha_k \end{bmatrix} \\ &= 1 + r_k^{\mathrm{T}} y_k + \alpha_k(\gamma_{k+1} + r_k^{\mathrm{T}} E_k y_k) \\ &= (1 - \alpha_k^2)\delta_k. \end{aligned} \tag{5.10}$$

因此，若在计算 α_k 时，利用(5.10)，便可将计算 α_k 的运算量减

少 $k-2$，从而就可将整个求解过程的运算量降为 n^2。

综合上面的讨论，可设计求解(5.2)的算法如下。

算法5.1

(1) 输入原始数据：$\gamma_1, \gamma_2, \cdots, \gamma_n$。

(2) $\zeta_1 := -\gamma_1$，$\delta := 1$，$\alpha := -\gamma_1$，$k := 1$。

(3) $\delta := (1-\alpha^2)\delta$，

$$\alpha := -\left(\gamma_{k+1} + \sum_{i=1}^{k} \gamma_{k-i+1}\zeta_i\right) \Big/ \delta,$$

$$\xi_i := \zeta_i + \alpha\zeta_{k-i+1}, \quad i = 1, 2, \cdots, k,$$

$$\zeta_i := \xi_i, \quad i = 1, 2, \cdots, k,$$

$$\zeta_{k+1} := \alpha.$$

(4) 如果 $k < n-1$，则 $k := k+1$，转步(3)；否则输出(5.2)之解 $y = (\zeta_1, \cdots, \zeta_n)^{\mathrm{T}}$，结束。

5.2 一般右端项的 Toeplitz 方程组

现在我们来考虑一般右端项的 Toeplitz 方程组

$$T_n x = b, \tag{5.11}$$

其中 $b = (\beta_1, \cdots, \beta_n)^{\mathrm{T}} \in \mathbb{R}^n$ 是已知向量。

类似于 Yule-Walker 方程组的求解过程，假定 $x_k \in \mathbb{R}^k$ 是 k 阶方程组

$$T_k x_k = (\beta_1, \cdots, \beta_k)^{\mathrm{T}} \quad (k = 1, 2, \cdots, n)$$

之解，则有

$$x_{k+1} = \begin{bmatrix} x_k + \mu_k E_k y_k \\ \mu_k \end{bmatrix}, \tag{5.12}$$

其中 y_k 是 k 阶 Yule-Walker 方程组(5.3)之解，E_k 是 k 阶反序单位阵，μ_k 由下式确定

$$\mu_k = (\beta_{k+1} - r_k^{\mathrm{T}} E_k x_k)/(1 + r_k^{\mathrm{T}} y_k), \tag{5.13}$$

这里 $r_k=(\gamma_1,\cdots,\gamma_k)^{\mathrm{T}}$。

这样，我们便可从 x_1 出发递推地求得(5.11)之解 x。这一求解过程可总结为如下算法。

算法5.2

(1) 输入原始数据：$\gamma_1,\cdots,\gamma_{n-1}$; $\beta_1,\cdots,\beta_n$。

(2) $\zeta_1:=-\gamma_1$，$\xi_1:=\beta_1$，$\delta:=1$，$\alpha:=-\gamma_1$，$k:=1$。

(3) $\delta:=(1-\alpha^2)\delta$,

$$\mu:=\left(\beta_{k+1}-\sum_{i=1}^{k}\gamma_i\xi_{k-i+1}\right)\Big/\delta,$$

$$\nu_i:=\xi_i+\mu\zeta_{k-i+1},\quad i=1,2,\cdots,k,$$

$$\xi_i:=\nu_i,\quad i=1,2,\cdots,k,$$

$$\xi_{k+1}:=\mu。$$

(4) 如果 $k<n-1$，则

$$\alpha:=-\left(\gamma_{k+1}+\sum_{i=1}^{k}\zeta_{k-i+1}\gamma_i\right)\Big/\delta,$$

$$\sigma_i:=\zeta_i+\alpha\zeta_{k-i+1},\qquad i=1,2,\cdots,k,$$

$$\zeta_i:=\sigma_i,\quad i=1,2,\cdots,k,$$

$$\zeta_{k+1}:=\alpha,$$

$k:=k+1$，转步(3)；

否则，输出 $x=(\xi_1,\cdots,\xi_n)^{\mathrm{T}}$，结束。

这一算法的运算量是 $2n^2$。

5.3 Toeplitz 矩阵的逆

最后，我们来考虑 T_n^{-1} 的计算问题。

设

$$T_n^{-1}=\begin{bmatrix} T_{n-1} & E_{n-1}r_{n-1} \\ r_{n-1}^{\mathrm{T}}E_{n-1} & 1 \end{bmatrix}^{-1}=\begin{bmatrix} X & v \\ v^{\mathrm{T}} & \sigma \end{bmatrix}\begin{matrix} n-1 \\ 1 \end{matrix},$$

$$\qquad\qquad\qquad\qquad\qquad\qquad\quad n-1\quad 1$$

其中 $r_{n-1}=(\gamma_1,\cdots,\gamma_{n-1})^{\mathrm{T}}$，$E_{n-1}$ 是 $n-1$ 阶反序单位矩阵。由

$$\begin{bmatrix} T_{n-1} & E_{n-1}r_{n-1} \\ r_{n-1}^{\mathrm{T}}E_{n-1} & 1 \end{bmatrix}\begin{bmatrix} X & v \\ v^{\mathrm{T}} & \sigma \end{bmatrix}=I_n=\begin{bmatrix} I_{n-1} & 0 \\ 0 & 1 \end{bmatrix},$$

可得

$$T_{n-1}X+E_{n-1}r_{n-1}v^{\mathrm{T}}=I_{n-1}, \tag{5.14}$$

$$T_{n-1}v+\sigma E_{n-1}r_{n-1}=0, \tag{5.15}$$

$$r_{n-1}^{\mathrm{T}}E_{n-1}v+\sigma=1。 \tag{5.16}$$

由(5.15)可得

$$v=\sigma E_{n-1}y_{n-1}, \tag{5.17}$$

其中 y_{n-1} 为 $n-1$ 阶 Yule-Walker 方程组的解。将(5.17)代入(5.16)，并整理，得

$$\sigma=1/(1+r_{n-1}^{\mathrm{T}}y_{n-1})。 \tag{5.18}$$

这样，我们只要求得 $n-1$ 阶 Yule-Walker 方程组之解 y_{n-1}，就可由(5.18)和(5.17)求出 T_n^{-1} 的最后一列和最后一行。

下面再来看 $X=[\xi_{ij}]$所具有的特性。从(5.14)可得

$$X=T_{n-1}^{-1}-T_{n-1}^{-1}E_{n-1}r_{n-1}v^{\mathrm{T}}=T_{n-1}^{-1}+vv^{\mathrm{T}}/\sigma, \tag{5.19}$$

其中的最后一个等式用到了 $T_{n-1}^{-1}E_{n-1}r_{n-1}=-E_{n-1}y_{n-1}$ 和(5.17)。由于 $T_{n-1}^{-1}=[t_{ij}]$是广对称的，故从(5.19)可得

$$\begin{aligned}\xi_{ij}&=t_{ij}+\nu_i\nu_j/\sigma\\&=t_{n-j,n-i}+\nu_i\nu_j/\sigma\\&=\xi_{n-j,n-i}+(\nu_i\nu_j-\nu_{n-i}\nu_{n-j})/\sigma,\end{aligned} \tag{5.20}$$

这里 ν_i 表示 v 的第 i 个分量。这也就是说，虽然X并非广对称的，但它的元素 ξ_{ij} 可由它的关于东北-西南对角线的对称元素 $\xi_{n-j,n-i}$ 确定。这样一来，我们就可利用 T_n^{-1} 的广对称性和(5.20)，从 T_n^{-1} 的边缘出发，逐层向内计算，求得 T_n^{-1} 的全部元素。为了清楚地了解这一过程，我们以$n=5$为例来说明其具体的计算过程。

设

$$T_5^{-1}=\begin{bmatrix} t_{11} & t_{12} & t_{13} & t_{14} & t_{15} \\ t_{21} & t_{22} & t_{23} & t_{24} & t_{25} \\ t_{31} & t_{32} & t_{33} & t_{34} & t_{35} \\ t_{41} & t_{42} & t_{43} & t_{44} & t_{45} \\ t_{51} & t_{52} & t_{53} & t_{54} & t_{55} \end{bmatrix}$$

第一步 先通过求解一个 4 阶 Yule-Walker 方程组得到 y_4，再利用(5.18)和(5.17)求出 T_5^{-1} 的最后一列和最后一行元素：$t_{15}=t_{51}$，$t_{25}=t_{52}$，$t_{35}=t_{53}$，$t_{45}=t_{54}$，t_{55}；然后再利用 T_5^{-1} 的广对称性，即知

$$t_{14}=t_{41}=t_{25},\quad t_{13}=t_{31}=t_{35},\quad t_{12}=t_{21}=t_{45},\quad t_{11}=t_{55}.$$

这样，我们就求得 T_5^{-1} 的最外一层的全部元素.

第二步 利用公式(5.20)，由第一步所得到的结果，可求得 $t_{24},t_{42},t_{34},t_{43}$ 和 t_{44}；然后再利用 T_5^{-1} 的广对称性即知 $t_{23}=t_{34}$，$t_{32}=t_{43},t_{22}=t_{44}$。于是，我们又求得 T_5^{-1} 的第二层的全部元素。

第三步 利用公式(5.20)，由第二步所得到的结果，可求得 t_{33}。至此，我们就已求得了 T_5^{-1} 的全部元素。

实际计算时，由于 T_n^{-1} 既是对称的又是广对称的，故只需计算 T_n^{-1} 的倒三角形部分的元素即可。如 $n=5$ 时仅需计算如下的 9 个元素即可：

$$\begin{matrix} t_{11} & t_{12} & t_{13} & t_{14} & t_{15} \\ & t_{22} & t_{23} & t_{24} & \\ & & t_{33} & & \end{matrix}$$

综上所述，我们可设计求 T_n^{-1} 的算法如下。

算法5.3

(1) 输入：$\gamma_1,\gamma_2,\cdots,\gamma_{n-1}$。

(2) 用算法 5.1 求解 $n-1$ 阶 Yule-Walker 方程组

$$T_{n-1}y=-(\gamma_1,\cdots,\gamma_{n-1})^{\mathrm{T}},$$

得 $y=(\zeta_1,\cdots,\zeta_{n-1})^{\mathrm{T}}$。

(3) $\sigma:=1\Big/\left(1+\sum_{k=1}^{n-1}\gamma_k\zeta_k\right)$,

$\nu_i:=\sigma\zeta_{n-i}\quad(i=1,2,\cdots,n-1)$,

$\xi_{11}:=\sigma$,

$\xi_{1j}:=\nu_{n-j+1}\quad(j=2,\cdots,n)$,

$i:=2$。

(4) $\xi_{ij}:=\xi_{i-1,j-1}+(\nu_{n-j+1}\nu_{n-i+1}-\nu_{i-1}\nu_{j-1})/\sigma$

$(j=i,i+1,\cdots,n-i+1)$。

(5) 如果 $i<\left[\dfrac{n-1}{2}\right]+1$，则 $i:=i+1$，转步(4)；否则输出有关信息，结束。

最后需指出的是，误差分析的结果表明：上述三种算法与应用 Cholesky分解来求解上述三类问题所引起的误差是差不多的。因此，这三种算法是数值稳定的。有关详细情况可参看文献[37]。

§6 条件数的估计和迭代改进

对于一个给定的线性方程组，我们应用某种数值方法求得它的一个近似解之后，一个很自然的问题就是：所得到的近似解的精确程度如何？如果精度太低，如何来改进它的精度呢？这一节，我们就以列选主元素的 Gauss 消去法为例来讨论这一问题。

6.1 条件数的估计

设我们应用列主元素的 Gauss 消去法在字长为 t 的10进制浮点数系下求解线性方程组

$$Ax=b。\tag{6.1}$$

则实际计算的经验表明，通常我们所得到的计算解 $\hat{x}$满足

$$(A+E)\hat{x}=b,\quad \frac{\|E\|_\infty}{\|A\|_\infty}\approx 10^{-t}. \tag{6.2}$$

需要注意的是，(6.2) 并不蕴含着所得到的计算解 $\hat{x}$ 已经是很精确的了。实际上，此时利用(1.6)可得如下估计

$$\frac{\|\hat{x}-x\|_\infty}{\|x\|_\infty}\approx 10^{-t}\kappa_\infty(A), \tag{6.3}$$

其中 x 表示方程组(6.1)的精确解。

由此可见，如果 $\kappa_\infty(A)\approx 10^s$，则由列主元素的 Gauss 消去法所产生的近似解 $\hat{x}$ 大约精确到小数点后 $t-s$ 位。

因此，我们要想对 $\hat{x}$ 的精度做出估计，就需估计出条件数的数量级。

由于 $\kappa_\infty(A)=\|A\|_\infty\|A^{-1}\|_\infty$，而

$$\|A\|_\infty=\max_{1\leq i\leq n}\sum_{j=1}^{n}|a_{ij}|$$

是容易计算的，故估计 $\kappa_\infty(A)$的主要困难在于对$\|A^{-1}\|_\infty$的估计。当然，我们可以借助于求解 n 个线性方程组

$$Ax_i=e_i,\quad i=1,2,\cdots,n,$$

而得到$A^{-1}=[x_1,x_2,\cdots,x_n]$；然后再求出$\|A^{-1}\|_\infty$。但这样做所需付出的工作量大约是求解 $\hat{x}$ 所需工作量的 3 倍之多，这当然是得不偿失的。

Cline, Moler, Stewart 和 Wilkinson(1979)(参见文献 [26])给出了一种非常实用的条件数的估计方法。这种方法在利用部分选主元素的 Gauss 消去法求得 $\hat{x}$ 的基础上，再增加较少的运算量，就可给出$\|A^{-1}\|_\infty$ 的数量级的较好的估计。其基本思想源于下面的基本事实：

$$Ay=d\ \Rightarrow\ \|A^{-1}\|_\infty\geqslant\|y\|_\infty/\|d\|_\infty.$$

如果能够选取向量 d 使$\|y\|_\infty/\|d\|_\infty$尽可能地靠近$\|A^{-1}\|_\infty$的话，我们就可用$\|y\|_\infty/\|d\|_\infty$来作为对$\|A^{-1}\|_\infty$的估计。易知，$\|y\|_\infty/\|d\|_\infty$

越大，估计就越精确。

那么，如何选取 d 可使 $\|y\|_\infty/\|d\|_\infty$ 尽可能地大呢？ A 的奇异值分解可为 d 的选取提供一些信息。

设 A 的奇异值分解为

$$A = U\Sigma V^{\mathrm{T}},$$

其中 $U=[u_1,\cdots,u_n]$ 和 $V=[v_1,\cdots,v_n]$ 是两个正交矩阵，$\Sigma=\mathrm{diag}(\sigma_1,\cdots,\sigma_n)$，$\sigma_1\geqslant\cdots\geqslant\sigma_n>0$。任取 $b\in\mathbb{R}^n$，考虑方程组

$$A^{\mathrm{T}}d=b \quad 和 \quad Ay=d。$$

将 b 表示为 $b=\sum_{i=1}^{n}\alpha_i v_i$，则易知上述二方程组的解分别为

$$d=\sum_{i=1}^{n}\frac{\alpha_i}{\sigma_i}u_i \quad 和 \quad y=\sum_{i=1}^{n}\frac{\alpha_i}{\sigma_i^2}v_i。$$

如果 A 是病态的，则一般应有其最小奇异值 σ_n 远远小于1。因此，只要 $|\alpha_n|$ 不要太小，就有

$$\|d\|_2\approx\frac{|\alpha_n|}{\sigma_n} \quad 和 \quad \|y\|_2\approx\frac{|\alpha_n|}{\sigma_n^2}。$$

从而有

$$\frac{\|y\|_2}{\|d\|_2}\approx\frac{1}{\sigma_n}=\|A^{-1}\|_2$$

应非常大。根据范数的等价性，进而有 $\|y\|_\infty/\|d\|_\infty$ 亦很大，且很接近于 $\|A^{-1}\|_\infty$。而选取 b 使 $|\alpha_n|$ 不要太小这样的条件是易于满足的，一般随机地选取即可。

基于上述讨论，在已利用部分选主元素的 Gauss 消去法求得分解 $PA=LU$ 的基础上，我们可以设计如下的条件数估计的基本方案：

（1）计算 $\|A\|_\infty$；

（2）对随机地选取的 b，求解 $U^{\mathrm{T}}x=b$；

（3）求解 $L^{\mathrm{T}}d=x$，$Lz=Pd$ 和 $Uy=z$；

（4） $\hat{\kappa}_\infty = \|A\|_\infty \|y\|_\infty / \|d\|_\infty$.

实现这一方案的关键在于应该如何具体地选取 b。要使 b 具有一定的随机性，通常可取 b 的分量的绝对值均为 1，即将 b 的各个分量等同的看待。这样剩下的问题就是 b 的每个分量的符号应该如何决定。

从前面的分析可知，b 在 v_n 上的投影越大越好；而 b 在 v_n 上的投影越大，应有 d 的范数也越大。因此，我们应该选取 b 使 $\|d\|_\infty$ 尽可能的大。从实际计算的经验知，一般 A 病态时，分解式 $PA = LU$ 中的 U 常是病态的，而 L 却是良态的。所以要使 $\|d\|_\infty$ 尽可能大，应选取 b 使 $U^{\mathrm{T}} x = b$ 之解 x 的无穷范数 $\|x\|_\infty$ 尽可能大。为此，我们来考察一下形如 $U^{\mathrm{T}} x = b$ 的方程组的求解过程。

设

$$U = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ & u_{22} & \cdots & u_{2n} \\ 0 & & \ddots & \vdots \\ & & & u_{nn} \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

则 $U^{\mathrm{T}} x = b$ 之解 x 可由如下算法求得：

（1） $\beta_i = 0$，$i = 1, 2, \cdots, n$；$k := 1$。

（2） $x_k := (b_k - \beta_k)/u_{kk}$，

$\beta_i := \beta_i + u_{ki} x_k$，$i = k+1, \cdots, n$。

（3） 如果 $k < n$，则 $k := k+1$，转步(2)；否则计算结束。

从这一算法可以看出，要使 $\|x\|_\infty$ 尽可能大的一种最简单的选取方法是取 $b_k = -\operatorname{sign} \beta_k$；另一种更可靠的方法是，$b_k$ 的符号不仅要使 $|x_k|$ 大，而且也要综合考虑受其影响的其他分量 x_i 的大小；而 x_i 的大小又与 $\beta_i + u_{ki} x_k$ 有关，因此我们先计算

$$x_k^+ = (1 - \beta_k)/u_{kk},$$

$$x_k^- = (-1 - \beta_k)/u_{kk},$$

$$s_k^+ = |x_k^+| + \sum_{i=k+1}^{n} |\beta_i + u_{ki} x_k^+|,$$

$$s_k^- = |x_k^-| + \sum_{i=k+1}^{n} |\beta_i + u_{ki} x_k^-|;$$

然后，如果 $s_k^+ \geqslant s_k^{-1}$，则选取 $b_k = 1$；否则选取 $b_k = -1$。

综合上面的讨论，我们就得到了实现基本方案中第二步的算法如下。

算法6.1

(1) 输入上三角阵 $U = [u_{ij}]$。

(2) $\beta_i := 0$，$i = 1, 2, \cdots, n$；$k := 1$。

(3) $x_k^+ := (1 - \beta_k)/u_{kk}$，$x_k^- := (-1 - \beta_k)/u_{kk}$，

$\beta_i^+ := \beta_i + u_{ki} x_k^+$，$i = k+1, \cdots, n$，

$\beta_i^- := \beta_i + u_{ki} x_k^-$，$i = k+1, \cdots, n$，

$$s_k^+ := |x_k^+| + \sum_{i=k+1}^{n} |\beta_i^+|,$$

$$s_k^- := |x_k^-| + \sum_{i=k+1}^{n} |\beta_i^-|.$$

(4) 如果 $s_k^+ \geqslant s_k^-$，则

$$x_k := x_k^+,\quad \beta_i := \beta_i^+,\quad i = k+1, \cdots, n;$$

否则

$$x_k := x_k^-,\qquad \beta_i := \beta_i^-,\qquad i = k+1, \cdots, n.$$

(5) 如果 $k < n$，则 $k := k+1$，转 步(3)；否则输出 x 结束。

由于基本方案中的其他三步都是很容易实现的，因此这里不再多述，建议读者作为练习，写出详细的条件数估计算法。

大家已经看到，这一方法的推理过程不是十分严格的，而且很多地方是依赖于实际计算经验的。但是大量的数值实验表明，在绝大多数情况下，它都可给出 $\kappa_\infty(A)$的数量级的很好估计，而且估计只需再增加 $O(n^2)$的运算量即可。因此，它是一种很有效的方法，其相应的程序已被收集在科学计算的程序库中。

6.2 迭代改进

如果我们已经判明计算解 $\hat{x}$ 的精度不满足要求，希望提高它的精度，则通常可使用下面的算法加以改进。

算法6.2

(1) $x:=\hat{x}$。

(2) $r:=b-Ax$（用双精度计算）。

(3) 求解 $Ly=Pr$，得 y。

(4) 求解 $Uz=y$，得 z。

(5) $x:=x+z$。

(6) 如果 x 已达到精度要求，则输出 x，结束；否则转步(2)。

注6.1 在上述迭代过程中，计算 $r=b-Ax$ 最好使用原始数据 A。因此，在对 A 进行分解时，最好将原来的 A 保留下来。

注6.2 实际计算经验表明：如果 A 病态的并不严重（$\varepsilon\kappa_\infty(A)<1$），那么利用算法 6.2 进行迭代，最终可产生达到机器精度的解；而如果 A 非常病态（$\varepsilon\kappa_\infty(A)\geqslant 1$），则一般用算法6.2并不能改进 x 的精度。

习　　题

1. 应用算法3.1于矩阵

$$A=\begin{bmatrix}0&4&1&2\\4&2&5&3\\1&5&0&1\\2&3&1&2\end{bmatrix}.$$

2. 用算法4.1求解 Vandermonde 方程组

$$V(1,2,3,4)z=(1,2,3,4)^{\mathrm{T}}.$$

3. 应用算法6.1估计矩阵

$$T=\begin{bmatrix}1 & 0 & x & -x\\ 0 & 1 & -x & x\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix},\quad x\in\mathbf{R}$$

的条件数 $\kappa_\infty(T)$。

4．考虑线性方程组 $Ax=b$，其中

$$A=\begin{bmatrix}2 & -1 & 1\\ -1 & 10^{-10} & 10^{-10}\\ 1 & 10^{-10} & 10^{-10}\end{bmatrix},\quad b=\begin{bmatrix}2(1+10^{-10})\\ -10^{-10}\\ 10^{-10}\end{bmatrix}.$$

(1) 验证 $x=(10^{-10},-1,1)$是方 程 组 的 解，且其条件数是 $\kappa_\infty(A)=2(10^{10}+1)\approx 2\times 10^{10}$。

(2) 证明：如果$|E|<10^{-8}|A|$，且$(A+E)y=b$，则有

$$|x-y|\leqslant 10^{-7}|x|.$$

这表明，即使A的条件数很大，A的元素的微小扰动未必一定会引起x的巨大变化。

(3) 定义 $D=\mathrm{diag}(10^{-5},10^5,10^5)$。证明：

$$\kappa_\infty(DAD)\leqslant 5.$$

5．设 $T_n\in\mathbf{R}^{n\times n}$ 是对称正定的Toeplitz 矩阵，$T_k(k=1,2,\cdots,n)$表示T_n的 k 阶顺序主子阵，设计一个计算 $\kappa_\infty(T_k)$的算法。

6．设计一个运算量为 $O(n^2)$的求解方程组 $Hx=b$ 的算法，其中H为上 Hessenberg 矩阵。

7．设计一个运算量为$O(n)$的算法来求解方程组 $Tx=b$，其中T是非奇异的三对角矩阵。

8．证明：

$$\|V(x_0,x_1,\cdots,x_n)^{-1}\|_\infty\leqslant\max_{0\leqslant k\leqslant n}\prod_{\substack{i=0\\ i\neq k}}^{n}\frac{1+|x_i|}{|x_k-x_i|}.$$

9. 假设 $P(A+E)=LU$，其中 P 是排列方阵，$L=[l_{ij}]$ 是满足 $|l_{ij}|\leqslant 1$ 的单位下三角矩阵，$U=[u_{ij}]$ 是上三角矩阵。证明

$$\kappa_{\infty}(A)\geqslant\frac{\|A\|_{\infty}}{\|E\|_{\infty}+\min\limits_{i}|u_{ii}|}.$$

第四章　线性方程组的迭代解法

§1　迭代法概述

第三章所介绍的方法，大多数均需对系数矩阵A进行分解，因而一般不能保持A的稀疏性。而实际应用中，特别是偏微分方程的数值求解时，常常遇到的恰恰就是大型稀疏线性方程组的求解问题。因此寻求能够保持稀疏性的有效解法就成为数值代数中一个非常重要的研究课题。

目前发展起来的方法主要有两类：一是充分利用所给矩阵A的特点，采用适当的主元素选取策略，使分解出的因子尽可能地保持稀疏性；二是迭代法。这一章我们将主要讨论古典迭代法，对第一类方法感兴趣的读者可参阅文献[58]。

迭代法一般可表述为：

$$x_k = \varphi_k(x_{k-1}, \cdots, x_{k-l}), \qquad k = l, l+1, \cdots, \qquad (1.1)$$

其中φ_k是从$\underbrace{\mathbb{R}^n \times \mathbb{R}^n \times \cdots \times \mathbb{R}^n}_{l\text{个}}$到$\mathbb{R}^n$上的算子，称作**迭代算子**，$x_0, \cdots, x_{l-1}$为**迭代初值**，可人为的给定。通常称迭代法(1.1)为l**步迭代法**；$l=1$时，亦称为单步迭代法。如果迭代算子φ_k与k无关，即$\varphi_k \equiv \varphi$，则称迭代法(1.1)为**定常迭代**；否则称之为**不定常迭代**。

这里，我们将着重讨论的是单步定常线性迭代法：

$$x_k = Gx_{k-1} + c, \qquad k = 1, 2, \cdots, \qquad (1.2)$$

其中$G \in \mathbb{R}^{n\times n}$称作**迭代矩阵**，$x_0$称为**初值**。

定义1.1　如果存在$x_* \in \mathbb{R}^n$，使得对任意的初值$x_0 \in \mathbb{R}^n$，由迭代法(1.2)产生的叙列$\{x_k\}_{k=1}^{\infty}$都收敛到x_*，即

$$\lim_{k\to\infty} x_k = x_*,$$

则称迭代法(1.2)是**收敛**的；否则称之为**发散**的。

如果迭代法(1.2)是收敛的，则必有

$$x_* = Gx_* + c. \tag{1.3}$$

现记

$$u_k = x_k - x_*, \tag{1.4}$$

则易证

$$u_k = G^k u_0. \tag{1.5}$$

由此可知，迭代法(1.2)收敛的充分必要条件是

$$\lim_{k\to\infty} G^k = 0.$$

再由第一章的定理3.8就可得如下关于迭代法(1.2)收敛性的判定定理。

定理1.1 迭代法(1.2)收敛的充分必要条件是

$$\rho(G) < 1. \tag{1.6}$$

在实际应用中，由于$\rho(G)$一般难以求出，我们通常并不利用$\rho(G)$是否小于 1 来判定迭代法(1.2) 是否收敛，而是利用一些易于计算的矩阵范数来判定。如果已求出$\|G\|$，只要$\|G\|<1$，则必有 $\rho(G)<1$，从而可断定迭代法(1.2)是收敛的。一般常用的矩阵范数是$\|\cdot\|_1$，$\|\cdot\|_\infty$和$\|\cdot\|_F$。这些范数都是用矩阵的元素表示的，因此用它们作为收敛的充分条件的判别标准是很方便的。

定理1.2 设$\|\cdot\|$是由向量范数$\|\cdot\|_a$诱导出的算子范数。如果$\|G\|<1$，则迭代法(1.2)收敛，且有

$$\|x_k - x_*\|_a < \frac{\|G\|^k}{1-\|G\|}\|x_0 - x_1\|_a \tag{1.7}$$

和

$$\|x_k - x_*\|_a < \frac{\|G\|}{1-\|G\|}\|x_k - x_{k-1}\|_a \tag{1.8}$$

对一切的自然数 k 成立。

证明留给读者。

从(1.8)可以看出：只要$\|G\|$不是很接近1，当相邻两次迭代向量x_k与x_{k-1}很接近时，则x_k与x_*也很接近。因此，常用量$\|x_k-x_{k-1}\|$是否已经适当小来判断迭代是否应当终止。但需特别注意的是，如果$\|G\|$很接近1，即使$\|x_k-x_{k-1}\|$已很小，也不能断定$\|x_k-x_*\|$很小。

对于一个收敛的迭代法(1.2)，其收敛的快慢是我们十分关心的问题，因此我们引进收敛速度的概念。

从(1.5)我们可以得到

$$\|u_k\|\leqslant\|G^k\|\|u_0\|,$$

于是，有

$$\frac{\|u_k\|}{\|u_0\|}\leqslant\|G^k\|. \tag{1.9}$$

上式表明$\|G^k\|$是迭代k次后误差与初始误差之比的上界。而平均地来说，每次迭代造成的误差的压缩率应正比于$\|G^k\|^{1/k}$。所以，我们有

定义1.2 迭代法(1.2)的**平均收敛速度**定义为

$$R_k(G)=-\frac{1}{k}\ln\|G^k\|. \tag{1.10}$$

从定义1.2不难看出，平均收敛速度$R_k(G)$不仅与迭代次数k有关，而且与所用的范数亦有关，这会在理论分析和实际应用中带来很大不便。为此，我们引进

定义1.3 迭代法(1.2)的**渐近收敛速度**定义为

$$R(G)=\lim_{k\to\infty}R_k(G)=-\ln\rho(G). \tag{1.11}$$

(1.11)的最后一个等号用了第一章的定理3.10。

§2 基本迭代法

设已给定线性方程组

$$Ax = b, \tag{2.1}$$

其中$A\in\mathbb{R}^{n\times n}$和$b\in\mathbb{R}^n$已知，$x\in\mathbb{R}^n$ 未知。现在我们来考虑如何通过构造形如(1.2)的迭代法来求(2.1)的解。

首先，我们自然希望构造出的迭代法(1.2)如果收敛，其极限就是方程组(2.1)的解。这就需要其迭代矩阵G和常向量c满足

$$QA = I - G \quad 和 \quad Qb = c, \tag{2.2}$$

其中$Q\in\mathbb{R}^{n\times n}$是某一非奇异矩阵。如果(2.2)成立，则称迭代法(1.2)与方程组(2.1)是**相容**的。从实用的目的考虑，当然我们只对相容的迭代法感兴趣。

现假定A有如下**分裂**

$$A = M - N, \tag{2.3}$$

其中M为非奇异矩阵。令$G = M^{-1}N$，$c = M^{-1}b$，则由此产生的迭代法(1.2)必然是相容的。此时，为了避免$M^{-1}N$和$M^{-1}b$的计算，可按如下方式进行迭代：

$$Mx_k = Nx_{k-1} + b, \qquad k = 1, 2, \cdots. \tag{2.4}$$

但这样一来，每次迭代就必须解一个系数矩阵为M的线性方程组。因此，我们自然希望M应具有某种特殊性，使这样的方程组易于求解；比如，M是对角形、上三角形、块对角形或块上三角形等。

基于这种思想，对A进行不同的分裂，就可构造出各种各样的相容的迭代法。下面我们就列举其中最基本的四种迭代法。

设A分块为

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1k} \\ A_{21} & A_{22} & \cdots & A_{2k} \\ \vdots & \vdots & & \vdots \\ A_{k1} & A_{k2} & \cdots & A_{kk} \end{bmatrix}, \tag{2.5}$$

其中$A_{ii}\in\mathbb{R}^{n_i\times n_i}$非奇异且系数阵为$A_{ii}$的线性方程组易于求解，$n_1 + n_2 + \cdots + n_k = n$。

令

$$D=\operatorname{diag}(A_{11},A_{22},\cdots,A_{kk}),\tag{2.6}$$

$$C_L=-\begin{bmatrix}0 & & & \\ A_{21} & 0 & & 0 \\ \vdots & \ddots & \ddots & \\ A_{k1} & \cdots & A_{k,k-1} & 0\end{bmatrix},$$

$$C_U=-\begin{bmatrix}0 & A_{12} & \cdots & A_{1k} \\ & \ddots & \ddots & \vdots \\ & & \ddots & A_{k-1,k} \\ 0 & & & 0\end{bmatrix},\tag{2.7}$$

$$L=D^{-1}C_L,\qquad U=D^{-1}C_U.\tag{2.8}$$

则

$$A=D-C_L-C_U=D(I-L-U).\tag{2.9}$$

1. Jacobi 迭代法(亦称简单迭代法)

A分裂为:

$$A=M_J-N_J,$$

其中

$$M_J=D,\qquad N_J=C_L+C_U.$$

迭代矩阵为:

$$\begin{aligned}J&=M_J^{-1}N_J=D^{-1}(C_L+C_U)\\&=L+U=I-D^{-1}A.\end{aligned}\tag{2.10}$$

迭代格式为:

$$Dx_m=(C_L+C_U)x_{m-1}+b,\qquad m=1,2,\cdots.\tag{2.11}$$

2. Gauss-Seidel 迭代法

A分裂为:

$$A=M_G-N_G,$$

其中

$$M_G=D-C_L,\qquad N_G=C_U.$$

迭代矩阵为:

$$\mathscr{L}_1=(D-C_L)^{-1}C_U=(I-L)^{-1}U. \tag{2.12}$$

迭代格式为:

$$(D-C_L)x_m=C_Ux_{m-1}+b, \quad m=1,2,\cdots. \tag{2.13}$$

3. 超松弛迭代法(简称 SOR 迭代法)

A 分裂为:

$$A=M_\omega-N_\omega,$$

其中

$$M_\omega=\frac{1}{\omega}D-C_L, \qquad N_\omega=\frac{1-\omega}{\omega}D+C_U,$$

ω 为非零实数，称作松弛因子。

迭代矩阵为:

$$\mathscr{L}_\omega=M_\omega^{-1}N_\omega=(I-\omega L)^{-1}(\omega U+(1-\omega)I). \tag{2.14}$$

迭代格式为:

$$(D-\omega C_L)x_m=(\omega C_U+(1-\omega)D)x_{m-1}+\omega b \quad (m=1,2,\cdots). \tag{2.15}$$

当$\omega=1$时，SOR迭代法就是 Gauss-Seidel 迭代法。因此，适当选择参数ω可望 SOR 迭代法比 Gauss-Seidel 迭代法具有更快的收敛速度。

4. 对称超松弛法(简称 SSOR 迭代法)

A 分裂为:

$$A=M_s-N_s,$$

其中

$$M_s=\frac{1}{\omega(2-\omega)}\{D-\omega(C_L+C_U)+\omega^2C_LD^{-1}C_U\}$$

$$=\frac{1}{\omega(2-\omega)}(D-\omega C_L)D^{-1}(D-\omega C_U),$$

$$N_s=\frac{1}{\omega(2-\omega)}\{(1-\omega)^2D+\omega(1-\omega)(C_L+C_U)+\omega^2C_LD^{-1}C_U\}$$

$$= \frac{1}{\omega(2-\omega)}[(1-\omega)D+\omega C_L]D^{-1}[(1-\omega)D+\omega C_U],$$

迭代矩阵为：

$$\mathscr{S}_\omega = M_s^{-1}N_s = \mathscr{U}_\omega \mathscr{L}_\omega,$$

其中

$$\mathscr{U}_\omega = (I-\omega U)^{-1}(\omega L+(1-\omega)I),$$

$\mathscr{L}_\omega$ 为 SOR 迭代矩阵。

迭代格式是：

$$(D-\omega C_L)x_{m-\frac{1}{2}} = \{\omega C_U+(1-\omega)D\}x_{m-1}+\omega b,$$

$$(D-\omega C_U)x_m = \{\omega C_L+(1-\omega)D\}x_{m-\frac{1}{2}}+\omega b.$$

由此可见，SSOR 迭代法实质上就是将 C_L 和 C_U 等同看待连续地使用两次 SOR 迭代。这样做的好处是：

（1）可充分利用内外存交换时得到的信息，减少内外存交换的次数，提高计算效率；

（2）在某些特殊问题中SOR 迭代法不收敛，但依然可构造出收敛的 SSOR 迭代法；

（3）SOR迭代法的渐近收敛速度对松弛因子 ω 的选择一般来说非常敏感，而SSOR 迭代法却不敏感。

注2.1 （1）在上面列举的四种基本迭代法中，当所有的 n_i 都等于 1 时，就是通常所讲的**点迭代**，如点Jacobi迭代，点SSOR迭代等。

（2）对于SSOR和SOR 迭代法而言，只有其松弛因子满足$0<\omega<2$时，才有可能收敛。因此，今后我们总假定ω满足这一条件。

§3 正定矩阵和某些迭代法的收敛性

迭代法构造出来之后，一个必须考虑的问题就是这一迭代法

是否收敛。这一节和下一节我们就来介绍几个关于古典迭代法的收敛性定理。

引理3.1 设$A\in\mathbb{C}^{n\times n}$是Hermite矩阵，且$A$分裂为$A=M-N$，其中$M$为非奇异矩阵。则$M^*+N$是Hermite矩阵，且对任意的$x\in\mathbb{C}^n$，有

$$x^*Ax-\tilde{x}^*A\tilde{x}=u^*(M^*+N)u, \tag{3.1}$$

其中$\tilde{x}=M^{-1}Nx$，$u=x-\tilde{x}$。

证明 由恒等式

$$\begin{aligned}M^*+N&=(A+N)^*+N=A+N^*+N\\&=M+N^*=(M^*+N)^*\end{aligned}$$

即知，M^*+N是Hermite矩阵，而且

$$M=M^*-N^*+N. \tag{3.2}$$

另一方面，由$M\tilde{x}=Nx$，可得

$$Mu=Mx-M\tilde{x}=Mx-Nx=Ax, \tag{3.3}$$

$$Nu=Nx-N\tilde{x}=M\tilde{x}-N\tilde{x}=A\tilde{x}. \tag{3.4}$$

从(3.3)和(3.4)，得

$$x^*Ax-\tilde{x}^*A\tilde{x}=x^*Mu-\tilde{x}^*Nu. \tag{3.5}$$

将(3.2)代入(3.5)，并注意到$x^*N^*=\tilde{x}^*M^*$，立即得(3.1)。

定理3.1 假设条件同引理3.1。则

(1) A和M^*+N正定$\Longrightarrow\rho(M^{-1}N)<1$；

(2) $\rho(M^{-1}N)<1$和M^*+N正定$\Longrightarrow A$正定。

证明 先证(1)。设$\lambda\in\lambda(M^{-1}N)$，即存在$x\in\mathbb{C}^n$，使得

$$M^{-1}Nx=\lambda x,\qquad x\neq 0. \tag{3.6}$$

对$x,\tilde{x}=\lambda x$和$u=(1-\lambda)x$应用等式(3.1)，得

$$(1-|\lambda|^2)x^*Ax=|1-\lambda|^2x^*(M^*+N)x. \tag{3.7}$$

此外，必有$\lambda\neq 1$；否则$Mx=Nx$，从而$Ax=0$，这必有$x=0$，这与$x\neq 0$的假设矛盾。于是，由A和M^*+N正定的假定，从(3.7)即知$|\lambda|<1$。注意到$\lambda\in\lambda(M^{-1}N)$的任意性，即有$\rho(M^{-1}N)<1$。

再证(2)。用反证法。若A不是正定的，则必存在$x_0\in\mathbf{C}^N$，使

$$\eta = x_0^* A x_0 < 0, \tag{3.8}$$

此处用到了$\rho(M^{-1}N)<1$蕴含着A非奇异。

从这一x_0出发，定义

$$x_k = M^{-1}N x_{k-1}, \quad k=1,2,\cdots,$$

则由$\rho(M^{-1}N)<1$的假定，知

$$\lim_{k\to\infty} x_k = \lim_{k\to\infty}(M^{-1}N)^k x_0 = 0. \tag{3.9}$$

现在令

$$u_k = x_k - x_{k-1}, \quad k=1,2,\cdots,$$

利用引理3.1，得

$$x_{k-1}^* A x_{k-1} - x_k^* A x_k = u_k^*(M^* + N)u_k \tag{3.10}$$

对一切自然数k都成立。

由x_0满足(3.8)可知$u_1\neq 0$。事实上，若$u_1=0$，则

$$-Ax_0 = Mu_1 = 0,$$

这与(3.8)矛盾。因此，对$k=1$应用(3.10)有

$$\begin{aligned} x_1^* A x_1 &= x_0^* A x_0 - u_1^*(M^* + N)u_1 \\ &< x_0^* A x_0 = \eta < 0, \end{aligned}$$

其中应用了M^*+N正定的假定。依此类推，可证

$$x_k^* A x_k < x_{k-1}^* A x_{k-1} \leqslant \eta < 0$$

对一切的自然数k成立。这与(3.9)矛盾。

定理3.2 设(2.5)所给的矩阵A是正定对称矩阵。则

(1) 当$2D-A$正定时，Jacobi迭代法收敛；

(2) 当$0<\omega<2$时，SOR和SSOR迭代法收敛。

证明 注意到，当A是实对称矩阵时，$D^*=D$，$C_L^*=C_U$，就有

$$M_J^* + N_J = D^* + C_L + C_U = D - A + D = 2D - A;$$

$$M_\omega^* + N_\omega = \left(\frac{1}{\omega}D - C_L\right)^* + \left(\frac{1-\omega}{\omega}D + C_U\right)$$

$$= \frac{2-\omega}{\omega}D;$$

$$M_s^* + N_s = \frac{1}{\omega(2-\omega)}[(D-\omega C_L)D^{-1}(D-\omega C_U)$$

$$+((1-\omega)D+\omega C_L)D^{-1}((1-\omega)D+\omega C_U)].$$

而A正定义蕴含着D正定和$(D-\omega C_L)^* = (D-\omega C_U)$非奇异，故$M^*+N_J$，$M_\omega^*+N_\omega$和$M_s^*+N_s$均是正定的。从而由定理3.3知，$\rho(J)=\rho(M_J^{-1}N_J)$，$\rho(\mathscr{L}_\omega)=\rho(M_\omega^{-1}N)$和$\rho(\mathscr{S}_\omega)=\rho(M_s^{-1}N_s)$均小于1。再应用定理1.1即知定理的结论成立。

注3.1 定理3.2的(2)蕴含着Gauss-Seidel迭代法也是收敛的(当然是在A正定的条件之下)。

定理3.3 设(2.5)所给的矩阵A是实对称的。则

(1) 当$2D-A$正定且Jacobi迭代法收敛时，A正定；

(2) 当D正定，且存在$\omega\in(0,2)$使得SOR或SSOR迭代法收敛时，A正定。

证明留作练习。

§4 H矩阵和某些迭代法的收敛性

定义4.1 设$A=[a_{ij}]\in\mathbb{R}^{n\times n}$。若$A$可表示为$A=sI-B$，其中$B\geqslant 0$，则当$s>\rho(B)$时，称$A$为**非奇异的$M$矩阵**，简称$M$矩阵；若$A$满足

$$a_{ij}\leqslant 0,\quad 1\leqslant i\neq j\leqslant n,$$
$$a_{ii}>0,\quad i=1,2,\cdots,n,$$

则称A为L**矩阵**。

定理4.1 $A=[a_{ij}]\in\mathbb{R}^{n\times n}$是$M$矩阵的充分必要条件是$A$是$L$矩阵，且$A^{-1}\geqslant 0$。

证明 *必要性* 设$A=sI-B$，$B\geqslant 0$，$s>\rho(B)$。由此易知

$a_{ij}\leqslant 0$，$i\neq j$，且 A 非奇异，并有

$$A^{-1}=(sI-B)^{-1}=\frac{1}{s}\left(I-\frac{1}{s}B\right)^{-1}=\frac{1}{s}\sum_{k=0}^{\infty}\left(\frac{1}{s}B\right)^{k}\geqslant 0.$$

记 $A^{-1}=[\beta_{ij}]$，由 $A^{-1}A=I$ 有

$$\sum_{k=1}^{n}\beta_{ik}a_{ki}=1,\quad i=1,2,\cdots,n.$$

于是

$$a_{ii}\beta_{ii}=1-\sum_{k\neq i}\beta_{ik}a_{ki}\geqslant 1,\quad i=1,2,\cdots,n.$$

而 $\beta_{ii}\geqslant 0$，故必有 $a_{ii}>0$，$i=1,2,\cdots,n$。因此 A 是 L 矩阵。

充分性　设 A 是 L 矩阵，且 $A^{-1}\geqslant 0$。由于 $a_{ii}>0$，现任取 $s>\max\limits_{i} a_{ii}$，则有 $B=sI-A\geqslant 0$。因此 $\rho(B)$ 是 B 的特征值，而且对应的特征向量可取作非负向量；即存在 $x\in\mathbb{R}^n$ 满足 $x\geqslant 0$，$x\neq 0$，使得

$$\rho(B)x=Bx=sx-Ax.$$

从而有

$$Ax=(s-\rho(B))x.$$

上式两边左乘 A^{-1}，得

$$x=(s-\rho(B))A^{-1}x.$$

由此即知，必有 $s>\rho(B)$。因此，A 是 M 矩阵。

定义4.2　设 $A=[a_{ij}]\in\mathbb{C}^{n\times n}$。令 $D=\mathrm{diag}(a_{11},\cdots,a_{nn})$，$C=D-A$。我们称矩阵 $|D|-|C|$ 为 A 的**比较矩阵**，记作 $\mathfrak{M}(A)$，即

$$\mathfrak{M}(A)=\begin{bmatrix}|a_{11}| & -|a_{12}| & \cdots & -|a_{1n}|\\ -|a_{21}| & |a_{22}| & \cdots & -|a_{2n}|\\ \cdots & \cdots & \cdots & \cdots\\ -|a_{n1}| & -|a_{n2}| & \cdots & |a_{nn}|\end{bmatrix}.$$

若 $\mathfrak{M}(A)$ 是非奇异的 M 矩阵，则称 A 为**非奇异的 H 矩阵**，简称 H

阵.

定义4.3 设 $A=[a_{ij}]\in\mathbb{C}^{n\times n}$。如果$A$满足

$$|a_{ii}|\geqslant\sum_{j=1}^{i-1}|a_{ij}|+\sum_{j=i+1}^{n}|a_{ij}|,\quad i=1,2,\cdots,n,$$

且至少有一个 i 使上述不等式严格成立，则称A为**弱严格对角占优矩阵**；如果上述n个不等式都严格成立，则称A为**严格对角占优矩阵**。

引理4.1 严格对角占优矩阵或不可分的弱严格对角占优矩阵是非奇异的。

证明留作练习。

定理4.2 严格对角占优矩阵或不可分的弱严格对角占优矩阵是H矩阵。

证明 设 $A\in\mathbb{C}^{n\times n}$ 是严格对角占优矩阵或不可分的弱严格对角占优矩阵。则由引理 4.1 知，$\mathfrak{M}(A)$必是非奇异的L矩阵。现任取 $s>\max\limits_{i}|a_{ii}|$，则

$$B=sI-\mathfrak{M}(A)\geqslant 0.$$

从而$\rho(B)$必是B的特征值。于是，据 Gerschgorin 圆盘定理，存在 i 使得

$$\rho(B)\leqslant|s-|a_{ii}||+\sum_{j\neq i}|a_{ij}|=s-\left(|a_{ii}|-\sum_{j\neq i}|a_{ij}|\right)\leqslant s.$$

但由于 $s-\rho(B)$是$\mathfrak{M}(A)$的特征值，而$\mathfrak{M}(A)$非奇异，从而 $s\neq\rho(B)$。因此，必有 $s>\rho(B)$，即A是H矩阵。

引理4.2 设 $M=[\omega_{ij}]\in\mathbb{C}^{n\times n}$ 是严格对角占优的。则对任意的 $N=[\eta_{ij}]\in\mathbb{C}^{n\times n}$ 有

$$\|M^{-1}N\|_\infty\leqslant\max_{1\leqslant i\leqslant n}\frac{\sum\limits_{j=1}^{n}|\eta_{ij}|}{|\omega_{ii}|-\sum\limits_{j\neq i}|\omega_{ij}|}.\tag{4.1}$$

证明 由算子范数的定义知，必存在 $y=(\beta_1,\cdots,\beta_n)^{\mathrm T}\in\mathbb{C}^n$，

$\|y\|_\infty=1$，使得

$$\|M^{-1}N\|_\infty=\|M^{-1}Ny\|_\infty.$$

现令 $x=(\alpha_1,\cdots,\alpha_n)^{\mathrm{T}}=M^{-1}Ny$,并假定 $|\alpha_{i_0}|=\max\limits_{1\le i\le n}|\alpha_i|=\|x\|_\infty$. 则由 $Mx=Ny$ 可得

$$\sum_{j=1}^{n}\omega_{i_0j}\alpha_j=\sum_{j=1}^{n}\eta_{i_0j}\beta_j.$$

于是，有

$$\begin{aligned}
&|\alpha_{i_0}|\Big(|\omega_{i_0i_0}|-\sum_{j\ne i_0}|\omega_{i_0j}|\Big)\\
&\leqslant|\omega_{i_0i_0}\alpha_{i_0}|-\sum_{j\ne i_0}|\omega_{i_0j}\alpha_j|\\
&\leqslant\Big|\sum_{j=1}^{n}\omega_{i_0j}\alpha_j\Big|=\Big|\sum_{j=1}^{n}\eta_{i_0j}\beta_j\Big|\\
&\leqslant\sum_{j=1}^{n}|\eta_{i_0j}\beta_j|\leqslant\sum_{j=1}^{n}|\eta_{i_0j}|.
\end{aligned}$$

从而，有

$$\|M^{-1}N\|_\infty=\|x\|_\infty=|\alpha_{i_0}|\leqslant\frac{\sum\limits_{j=1}^{n}|\eta_{i_0j}|}{|\omega_{i_0i_0}|-\sum\limits_{j\ne i_0}|\omega_{i_0j}|}.$$

由此立知（4.1）成立.

引理4.3 设 $M=[\omega_{ij}]\in\mathbb{R}^{n\times n}$ 是严格对角占优的L矩阵，$N=[\eta_{ij}]\in\mathbb{R}^{n\times n}$ 是非负的. 则

$$\rho(M^{-1}N)\geqslant\min_{1\le i\le n}\frac{\sum\limits_{j=1}^{n}|\eta_{ij}|}{|\omega_{ii}|-\sum\limits_{j\ne i}|\omega_{ij}|}. \tag{4.2}$$

证明 令 $y=(\beta_1,\cdots,\beta_n)^T=M^{-1}Ne$，其中 $e=(1,1,\cdots,1)^T$. 则由假定知 $y\geqslant 0$。设 $\beta_{i_0}=\min\limits_{1\leqslant i\leqslant n}\beta_i$. 由 $My=Ne$ 得

$$\sum_{j=1}^{n}\eta_{i_0 j}=\sum_{j=1}^{n}\omega_{i_0 j}\beta_j=\omega_{i_0 i_0}\beta_{i_0}-\sum_{j\neq i_0}|\omega_{i_0 j}|\beta_j$$

$$\leqslant\beta_{i_0}\Big(\omega_{i_0 i_0}-\sum_{j\neq i_0}|\omega_{i_0 j}|\Big).$$

因此，记 $\beta_{ij}=e_i^T M^{-1}Ne_j$，则有

$$\rho(M^{-1}N)\geqslant\min_{1\leqslant i\leqslant n}\sum_{j=1}^{n}\beta_{ij}=\min_{1\leqslant i\leqslant n}\beta_i=\beta_{i_0}$$

$$\geqslant\sum_{i=1}^{n}\eta_{i_0 j}\Big/\Big(\omega_{i_0 i_0}-\sum_{j\neq i_0}|\omega_{i_0 j}|\Big),$$

其中的第一个不等式用到了非负矩阵的一条性质（详见本章习题8)和假设条件蕴含着 $M^{-1}\geqslant 0$。由上述不等式立即知(4.2)成立。

有了前面的准备工作，我们现在可以给出一类更广的迭代法 AOR 和 SAOR 的收敛定理。

对于给定的 $A=[a_{ij}]\in\mathbb{R}^{n\times n}$，记

$$D=\mathrm{diag}(a_{11},\cdots,a_{nn}),$$

$$C_L=-\begin{bmatrix}0 & & & & \\ a_{21} & 0 & & 0 & \\ \vdots & \ddots & \ddots & & \\ \vdots & & \ddots & \ddots & \\ a_{n1} & \cdots & \cdots & a_{n,n-1} & 0\end{bmatrix},$$

$$C_U=-\begin{bmatrix}0 & a_{12} & \cdots & \cdots & a_{1n}\\ & 0 & \ddots & & \vdots\\ & & \ddots & \ddots & \vdots\\ & 0 & & \ddots & a_{n-1,n}\\ & & & & 0\end{bmatrix},$$

即将A分裂为$A=D-C_L-C_U$。

AOR 迭代法的迭代矩阵为

$$\mathscr{L}_{r,\omega}(A)=(D-rC_L)^{-1}[(1-\omega)D+(\omega-r)C_L+\omega C_U], \tag{4.3}$$

其中r和ω为松弛参数。这种迭代法是由 Hadjidimas 于1978年提出的，通常称作**快速松弛迭代法**，简称 AOR 迭代法。易知：

当$r=\omega$时，AOR 法即为 SOR 迭代法；

当$r=\omega=1$时，AOR 法即为 Gauss-Seidel 迭代法；

当$r=0$，$\omega=1$时，AOR 法即为 Jacobi 迭代法。

SAOR 迭代法的迭代矩阵为

$$\mathscr{S}_{r,\omega}(A)=\mathscr{U}_{r,\omega}(A)\mathscr{L}_{r,\omega}(A), \tag{4.4}$$

其中$\mathscr{L}_{r,\omega}(A)$由(4.3)给出，$\mathscr{U}_{r,\omega}(A)$是将(4.3)中的$C_L$和$C_U$互换而得到的，即

$$\mathscr{U}_{r,\omega}(A)=(D-rC_U)^{-1}[(1-\omega)D+(\omega-r)C_U+\omega C_L].$$

显然$\mathscr{S}_{r,\omega}(A)$是 SSOR 迭代矩阵$\mathscr{S}_\omega$的推广。通常称$\mathscr{S}_{r,\omega}(A)$为对称 AOR 迭代矩阵，相应的迭代法为**对称快速松弛法**，简称为 SAOR 迭代法。

此外，记

$$\Omega(A)=\{B=[\beta_{ij}]\in\mathbb{R}^{n\times n}\mid|\beta_{ij}|=|\alpha_{ij}|,\ 1\leqslant i,j\leqslant n\},$$

称作A的**等模矩阵集合**。

定理4.3 设$A=[\alpha_{ij}]\in\mathbb{R}^{n\times n}$满足$\alpha_{ii}\neq0$，$i=1,\cdots,n$。则$A$是$H$矩阵的必要充分条件是$\rho(|J|)<1$，其中$J$表示对应于$A$的点 Jacobi 迭代矩阵。

证明 记$D=\operatorname{diag}(\alpha_{11},\cdots,\alpha_{nn})$，$C=D-A$。则$J=D^{-1}C$，

$$\mathfrak{M}(A)=|D|-|C|=|D|(I-|J|).$$

如果$\rho(|J|)<1$，则$(I-|J|)^{-1}$存在，且

$$(I-|J|)^{-1}=\sum_{k=0}^{\infty}|J|^k\geqslant0.$$

因而$\mathfrak{M}(A)$可逆，且

$$\mathfrak{M}(A)^{-1}=(I-|J|)^{-1}|D|^{-1}\geqslant 0.$$

即A是H矩阵。

反之，若A为H矩阵，则$\mathfrak{M}(A)^{-1}\geqslant 0$。从而$(I-|J|)$可逆，且有

$$(I-|J|)^{-1}=\mathfrak{M}(A)^{-1}|D|\geqslant 0.$$

再注意到

$$(I-|J|)(I+|J|+\cdots+|J|^k)=I-|J|^{k+1},$$

便有

$$\sum_{i=0}^{k}|J|^i=(I-|J|)^{-1}(I-|J|^{k+1})\leqslant(I-|J|)^{-1},$$

即$\sum_{i=0}^{k}|J|^i$对一切的k有上界$(I-|J|)^{-1}$。因此$\sum_{k=0}^{\infty}|J|^k$收敛。故必有$\rho(|J|)<1$。

定理4.4 设$A=[a_{ij}]\in\mathbb{R}^{n\times n}$满足$a_{ii}\neq 0$，$i=1,2,\cdots,n$，$0\leqslant r\leqslant\omega$。则下列三条等价：

(1) A是H矩阵；

(2) 对任意的$G\in\Omega(A)$和任意的$\omega\in(0,2/(1+\rho(|J|))$，有$\rho(\mathscr{L}_{r,\omega}(G))<1$；

(3) 对任意的$G\in\Omega(A)$和任意的$\omega\in(0,2/(\rho(|J|)+1))$，有$\rho(\mathscr{S}_{r,\omega}(G))<1$。

这里的J如定理4.3所述。

证明 （Ⅰ）先考虑A不可分的情形。

我们先证(1)成立蕴含着(2)和(3)成立。此时，不妨就假定$G=A$。由A不可分知$|J|=|D^{-1}||C|$是非负不可分矩阵。故由Perron-Frobenius定理知，存在$x=(\alpha_1,\cdots,\alpha_n)^{\mathrm{T}}\in\mathbb{R}^n$，$x>0$，满足

$$|D|^{-1}|C|x=\rho(|J|)x. \tag{4.5}$$

现在令

$$Q=\mathrm{diag}(\alpha_1,\cdots,\alpha_n),\quad \tilde{A}=[\tilde{a}_{ij}]=AQ.$$

则从(4.5)可得

$$\sum_{j\neq i}|\tilde{a}_{ij}|=\rho(|J|)|\tilde{a}_{ii}|,\quad i=1,2,\cdots,n, \tag{4.6}$$

而且易知

$$Q^{-1}\mathscr{L}_{r,\omega}(A)Q=\mathscr{L}_{r,\omega}(\tilde{A})=\tilde{M}^{-1}\tilde{N}, \tag{4.7}$$

其中

$$\tilde{M}=\tilde{D}-r\tilde{C}_L,$$

$$\tilde{N}=(1-\omega)\tilde{D}+(\omega-r)\tilde{C}_L+\omega\tilde{C}_U,$$

这里 $\tilde{D}=DQ$，$\tilde{C}_L=C_LQ$，$\tilde{C}_U=C_UQ$。

下面证明 $\tilde{M}$ 是严格对角占优的。由于 $0\leqslant r<\dfrac{2}{1+\rho(|J|)}$，$\rho(|J|)<1$，故从(4.6)可得

$$\begin{aligned}
|\tilde{a}_{ii}|-r\sum_{j=1}^{i-1}|\tilde{a}_{ij}| &\geqslant |\tilde{a}_{ii}|-\frac{2}{1+\rho(|J|)}\sum_{j=1}^{i-1}|\tilde{a}_{ij}| \\
&=\frac{1}{1+\rho(|J|)}\left[(1+\rho(|J|))|\tilde{a}_{ii}|-2\cdot\sum_{j=1}^{i-1}|\tilde{a}_{ij}|\right] \\
&>\frac{1}{1+\rho(|J|)}\left[2\sum_{j\neq i}|\tilde{a}_{ij}|-2\sum_{j=1}^{i-1}|\tilde{a}_{ij}|\right] \\
&=\frac{2}{1+\rho(|J|)}\sum_{j=i+1}^{n}|\tilde{a}_{ij}|\geqslant 0,\quad i=1,\cdots,n.
\end{aligned}$$

因而 $\tilde{M}$ 是严格对角占优矩阵。利用引理 4.2 可得

$$\rho(\mathscr{L}_{r,\omega}(A))=\rho(\mathscr{L}_{r,\omega}(\tilde{A}))\leqslant\|\mathscr{L}_{r,\omega}(\tilde{A})\|_\infty=\|\tilde{M}^{-1}\tilde{N}\|_\infty$$

$$\leqslant\max_{1\leqslant i\leqslant n}\frac{|1-\omega||\tilde{a}_{ii}|+(\omega-r)\sum\limits_{j=1}^{i-1}|\tilde{a}_{ij}|+\omega\sum\limits_{j=i+1}^{n}|\tilde{a}_{ij}|}{|\tilde{a}_{ii}|-r\sum\limits_{j=1}^{i-1}|\tilde{a}_{ii}|}$$

$$= \max_{1 \leqslant i \leqslant n} \frac{[|1-\omega| + \omega\rho(|J|)]|\tilde{a}_{ii}| - r\sum_{j=1}^{i-1}|\tilde{a}_{ij}|}{|\tilde{a}_{ii}| - r\sum_{j=1}^{i-1}|\tilde{a}_{ij}|}$$

$$\leqslant |1-\omega| + \omega\rho(|J|) \tag{4.8}$$

$$< 1. \tag{4.9}$$

其中用到了数值不等式：若正数 a,b,c 满足 $a-c>0$ 和 $a \geqslant b$，则 $(b-c)/(a-c) \leqslant b/a$；以及假定 $0<\omega<2/[1+\rho(|J|)]$蕴含着 $|1-\omega| + \omega\rho(|J|) < 1$。

同理可证

$$\|\mathscr{U}_{r,\omega}(\tilde{A})\|_\infty \leqslant |1-\omega| + \omega\rho(|J|) < 1. \tag{4.10}$$

再注意到$\mathscr{S}_{r,\omega}(\tilde{A}) = Q^{-1}\mathscr{S}_{r,\omega}(A)Q$，从(4.8)和(4.10)即得

$$\begin{aligned}\rho(\mathscr{S}_{r,\omega}(A)) &= \rho(\mathscr{S}_{r,\omega}(\tilde{A})) \leqslant \|\mathscr{S}_{r,\omega}(\tilde{A})\|_\infty \\ &= \|\mathscr{U}_{r,\omega}(\tilde{A})\mathscr{L}_{r,\omega}(\tilde{A})\|_\infty \\ &\leqslant [|1-\omega| + \omega\rho(|J|)]^2 < 1. \end{aligned} \tag{4.11}$$

再证(1)不成立必有(2)和(3)不成立。这只需证，在 A 非 H 矩阵的条件下，必存在矩阵 $G \in \Omega(A)$ 和 r,ω 满足 $0 \leqslant r \leqslant \omega < 2/[1+\rho(|J|)]$，使得

$$\rho(\mathscr{L}_{r,\omega}(G)) \geqslant 1 \text{ 和 } \rho(\mathscr{S}_{r,\omega}(G)) \geqslant 1$$

成立即可。

现就取 $G = \mathfrak{M}(A)$。则 G 是 L 矩阵但不是 H 矩阵。于是据定理 4.3 知必有 $\rho(|J|) \geqslant 1$。因此，从 (4.6) 可得

$$|\tilde{a}_{ii}| \leqslant \sum_{j \neq i}|\tilde{a}_{ij}|, \quad i = 1,2,\cdots,n. \tag{4.12}$$

由于$|\tilde{a}_{ii}|>0$，$i=1,2,\cdots,n$，故可取 r 为充分小的正数，使得

$$|\tilde{a}_{ii}| - r\sum_{j=1}^{i-1}|\tilde{a}_{ij}| > 0, \quad |\tilde{a}_{ii}| - r\sum_{j=i+1}^{n}|\tilde{a}_{ij}| > 0,$$
$$i = 1,2,\cdots,n. \tag{4.13}$$

这样，对任意的$\omega \in (r, 2/[1+\rho(|J|)])$，应用引理4.3于$M=|\tilde{D}|-r|\tilde{C}_L|$和$N=(1-\omega)|\tilde{D}|+(\omega-r)|\tilde{C}_L|+\omega|\tilde{C}_U|$上，可得

$$\rho(\mathscr{L}_{r,\omega}(\mathfrak{M}(A)))=\rho(\mathscr{L}_{r,\omega}(\mathfrak{M}(\tilde{A})))=\rho(M^{-1}N)$$

$$\geqslant \min_{1\leq i\leq n}\sum_{j=1}^{n} l_{ij}$$

$$\geqslant \min_{1\leq i\leq n}\frac{(1-\omega)|\tilde{a}_{ii}|+(\omega-r)\sum_{j=1}^{i-1}|\tilde{a}_{ij}|+\omega\sum_{j=i+1}^{n}|\tilde{a}_{ij}|}{|\tilde{a}_{ii}|-r\sum_{j=1}^{i-1}|\tilde{a}_{ij}|}$$

$$\geqslant 1, \tag{4.14}$$

其中$l_{ij}=e_i^{\mathrm{T}}M^{-1}Ne_j=e_i^{\mathrm{T}}\mathscr{L}_{r,\omega}(\mathfrak{M}(\tilde{A}))e_j$，$i,j=1,2,\cdots,n$。

同理可证

$$\min_{1\leq i\leq n}\sum_{j=1}^{n}u_{ij}\geqslant 1, \tag{4.15}$$

其中$u_{ij}=e_i^{\mathrm{T}}\mathscr{U}_{r,\omega}(\mathfrak{M}(\tilde{A}))e_j$，$i,j=1,2,\cdots,n$。

记

$$\beta_{ij}=e_i^{\mathrm{T}}(\mathscr{U}_{r,\omega}(\mathfrak{M}(\tilde{A}))\mathscr{L}_{r,\omega}(\mathfrak{M}(\tilde{A})))e_j,\quad i,j=1,2,\cdots,n.$$

则有

$$\min_{1\leq i\leq n}\sum_{j=1}^{n}\beta_{ij}=\min_{1\leq i\leq n}\sum_{j=1}^{n}\sum_{k=1}^{n}u_{ik}l_{kj}$$

$$=\min_{1\leq i\leq n}\sum_{k=1}^{n}u_{ik}\sum_{j=1}^{n}l_{kj}$$

$$\geqslant \min_{1\leq i\leq n}\sum_{k=1}^{n}u_{ik}\left(\min_{1\leq k\leq n}\sum_{j=1}^{n}l_{kj}\right)$$

$$=\left(\min_{1\leq i\leq n}\sum_{k=1}^{n}u_{ik}\right)\left(\min_{1\leq k\leq n}\sum_{j=1}^{n}l_{kj}\right)\geqslant 1,$$

最后一个不等式应用了（4.14）和（4.15）；从而

$$\rho(\mathscr{S}_{r,\omega}(\mathfrak{M}(A))=\rho(\mathscr{S}_{r,\omega}(\mathfrak{M}(\tilde{A})))\geqslant \min_{1\leq i\leq n}\sum_{j=1}^{n}\beta_{ij}\geqslant 1,$$

这样，我们就在A不可分的条件之下证明了定理成立。

（Ⅱ）考虑A可分的情形。

此时亦先证(1)蕴含着(2)和(3)．仍不妨假定$G=A$．现将A所有等于零的元素都换成ε而得矩阵A_ε，则A_ε是不可分的。由于A是H矩阵，故$\rho(|J|)<1$．记A_ε对应的Jacobi迭代矩阵为J_ε，则$J_\varepsilon=D^{-1}C_\varepsilon$．(因$A$的对角元素均不为零，故$A$与$A_\varepsilon$有相同的对角元素)．因此，当$\varepsilon\to0$时，有$\rho(|J_\varepsilon|)\to\rho(|J|)<1$．所以，必存在$\varepsilon_1>0$，使当$0<\varepsilon<\varepsilon_1$时，有$\rho(|J_\varepsilon|)<1$，从而有$A_\varepsilon$是$H$矩阵．应用(Ⅰ)所证不等式(4.8)和(4.11)知，当$0<\varepsilon<\varepsilon_1$时，有

$$\rho(\mathscr{L}_{r,\omega}(A_\varepsilon))\leqslant|1-\omega|+\omega\rho(|J_\varepsilon|) \qquad (4.16)$$

和

$$\rho(\mathscr{S}_{r,\omega}(A_\varepsilon))\leqslant[|1-\omega|+\omega\rho(|J_\varepsilon|)]^2 \qquad (4.17)$$

对一切的r和ω满足$0\leqslant r\leqslant\omega<2/[1+\rho(|J_\varepsilon|)]$成立。

现任意取定r和ω满足$0\leqslant r\leqslant\omega<2/[1+\rho(|J|)]$，则必存在正数$\varepsilon_2\leqslant\varepsilon_1$，使当$0<\varepsilon<\varepsilon_2$时，有

$$0\leqslant r\leqslant\omega<\frac{2}{[1+\rho(|J_\varepsilon|)]}.$$

进而，当$0<\varepsilon<\varepsilon_2$时，有(4.16)和(4.17)成立，故令$\varepsilon\to0$，即得

$$\rho(\mathscr{L}_{r,\omega}(A))\leqslant|1-\omega|+\omega\rho(|J|)<1,$$
$$\rho(\mathscr{S}_{r,\omega}(A))\leqslant[|1-\omega|+\omega\rho(|J|)]^2<1.$$

这说明(2)和(3)成立。

用类似的技巧可证(1)不成立时，亦有(2)和(3)不成立．详细证明留作练习读者自己补出。

作为这一定理的一个直接推论，我们有

推论4.1 如果A是严格对角占优矩阵或不可分弱严格对角占优矩阵，$0\leqslant\omega<2/[1+\rho(|J|)]$，则对应的点Jacobi，点Gauss-

Seidel，点 SOR 和点 SSOR 迭代法都是收敛的。

§5 多项式加速

从前面两节的收敛性定理可以看出，用古典迭代法求解线性方程组时，一般收敛性是难以保证的。此外，即使收敛，也可能收敛的很慢。因此，我们自然希望能找到一种方法，在不增加太多的运算量的前提下，对原迭代产生的序列加以适当的修正，使得原来不收敛的变得收敛，原来收敛的收敛得更快，这就是本节将要讨论的中心问题。

经典分析中求和法的基本思想为我们寻求这样的方法奠定了理论基础。设$\{x_k\}_{k=0}^{\infty}$是一给定的实数序列。最简单的求和法是构造新序列

$$y_k=\frac{1}{k+1}(x_0+x_1+\cdots+x_k),\quad k=0,1,\cdots. \tag{5.1}$$

这样做的结果是：若$\{x_k\}_{k=0}^{\infty}$收敛，则$\{y_k\}_{k=0}^{\infty}$亦收敛，而且它们的极限相同；若$\{x_k\}_{k=0}^{\infty}$不收敛，则$\{y_k\}_{k=0}^{\infty}$仍可能收敛。更一般的求和法是，先选定一组参数

$$\begin{array}{llll} a_{00} & & & \\ a_{10} & a_{11} & & \\ \cdots\cdots & & & \\ a_{k0} & a_{k1} & \cdots & a_{kk} \\ \cdots\cdots & & & \end{array} \tag{5.2}$$

满足

$$\sum_{i=0}^{k}a_{ki}=1,\quad k=0,1,\cdots; \tag{5.3}$$

然后，构造新的序列

$$y_k=\sum_{j=0}^{k}a_{kj}x_j,\quad k=0,1,\cdots. \tag{5.4}$$

当 $a_{ki}=0(i=0,1,\cdots,k-1)$，$a_{kk}=1$ 时，$y_k=x_k$；当 $a_{ki}=1/(k+1)$时，(5.3)就是前面所讲的简单求和法。由于这里的 a_{ki} 有较大的选择自由度，因此可望通过 a_{ki} 的适当选取使新的序列 $\{y_k\}_{k=1}^{\infty}$ 具有更好的收敛性。

现在我们来考虑如何把求和法的基本思想应用到我们所关心的问题上。假定 $x_0,x_1,\cdots$是由与方程组(2.1)相容的迭代法

$$x_{k+1}=Gx_k+c,\quad k=0,1,\cdots$$

产生的向量序列。对于给定的参数 (5.2)，类比于(5.4)，亦构造新的序列

$$\tilde{x}_k=\sum_{i=0}^{k}a_{kj}x_j,\qquad k=0,1,\cdots. \tag{5.5}$$

令

$$u_k=\tilde{x}_k-x_*,\quad k=0,1,\cdots, \tag{5.6}$$

其中 x_* 满足 $x_*=Gx_*+c$。自然，我们希望通过参数的适当选取使 u_k 尽可能快地收敛到零。将(5.5)代入(5.6)即得

$$\begin{aligned}u_k&=\sum_{j=0}^{k}a_{kj}x_j-x_*=\sum_{j=0}^{k}a_{kj}(x_j-x_*)\\&=\sum_{j=0}^{k}a_{kj}G^j(x_0-x_*)\\&=\left(\sum_{j=0}^{k}a_{kj}G^j\right)u_0.\end{aligned} \tag{5.7}$$

记

$$q_k(t)=\sum_{j=0}^{k}a_{kj}t^j,$$

则 (5.7)可写作

$$u_k=q_k(G)u_0. \tag{5.8}$$

因而，欲使 u_k 尽可能快地收敛到零，自然就应选取参数 a_{kj} 使 $q_k(G)$的谱半径尽可能的小，这等价于求解如下的优化问题：求 $q_k\in\mathscr{P}_k^{(1)}$，使得

$$\max_{\lambda\in\lambda(G)}|q_k(\lambda)| = \min_{p_k\in\mathscr{P}_k^{(1)}}\max_{\lambda\in\lambda(G)}|p_k(\lambda)|.$$

其中 $\mathscr{P}_k^{(1)}$ 是次数不超过 k 且满足 $p_k(1)=1$ 的实系数多项式 $p_k(t)$ 的全体。

然而，求解这样一个优化问题并非一件易事。这是因为我们一般很难详细知道 G 的所有特征值，即便知道了 G 的所有特征值，求解离散集合上的一个最佳一致逼近问题，理论上亦有一定的困难。所以，我们转而考虑一个连续集上的最佳一致逼近问题，即求 $q_k\in\mathscr{P}_k^{(1)}$，使得

$$\max_{\mu\in\bar{S}_G}|q_k(\mu)| = \min_{p_k\in\mathscr{P}_k^{(1)}}\max_{\mu\in\bar{S}_G}|p_k(x)|, \tag{5.9}$$

其中 $\bar{S}_G$ 为包含 $\lambda(G)$ 的某一凸集。显然，$\bar{S}_G$ 取的越小越好，而且可根据 G 的特点进行估计。下面我们分三种情况来讨论上述优化问题。

1. 设 G 的特征值都是实数，并满足

$$\lambda(G)\subset[\alpha,\beta]\subset(-\infty,1).$$

此时，取 $\bar{S}_G=[\alpha,\beta]$，则(5.9)就成为：求 $q_k\in\mathscr{P}_k^{(1)}$，使得

$$\max_{t\in[\alpha,\beta]}|q_k(t)| = \min_{p_k\in\mathscr{P}_k^{(1)}}\max_{t\in[\alpha,\beta]}|p_k(t)|. \tag{5.10}$$

著名的 Chebyshev 定理（参见文献[11]）告诉我们，优化问题(5.10)有唯一的解

$$q_k(t) = T_k\left(\frac{2t-\beta-\alpha}{\beta-\alpha}\right)\Big/T_k\left(\frac{2-\beta-\alpha}{\beta-\alpha}\right), \tag{5.11}$$

其中 $T_k(t)$ 是 k 阶 Chebyshev 多项式，由下面的递推公式给出：

$$T_0(t)\equiv 1,\quad T_1(t)=t,$$
$$T_{k+1}(t)=2tT_k(t)-T_{k-1}(t),\quad k=2,3,\cdots. \tag{5.12}$$

这样，在理论上，我们已经找到了在上述意义下的最佳参数 α_{kj}。但实际使用时，这是行不通的。这里的原因有二：其一是整个迭代过程需要保存所有的 $x_1,\cdots,x_k$，当 k 较大时，要占据太

多的内存；其二是求出 q_k 的系数也是很费时间的．幸运的是，利用 Chebyshev 多项式的三项递推公式，我们能够完全避免上述两个问题的出现，而给出直接求 $\tilde{x}_k$ 的非常简单的迭代公式．

记

$$\xi = \frac{2-\beta-\alpha}{\beta-\alpha}, \quad l(t) = \frac{2t-\beta-\alpha}{\beta-\alpha}.$$

则

$$q_{k+1}(t) = \frac{T_{k+1}(l(t))}{T_{k+1}(\xi)} = \frac{2l(t)T_k(l(t)) - T_{k-1}(l(t))}{T_{k+1}(\xi)}$$

$$= \frac{2T_k(\xi)}{T_{k+1}(\xi)} \cdot l(t) \cdot \frac{T_k(l(t))}{T_k(\xi)} - \frac{T_{k-1}(\xi)}{T_{k+1}(\xi)} \cdot \frac{T_{k-1}(l(t))}{T_{k-1}(\xi)}$$

$$= \frac{2T_k(\xi)}{T_{k+1}(\xi)} l(t) q_k(t) - \frac{T_{k-1}(\xi)}{T_{k+1}(\xi)} q_{k-1}(t).$$

于是有

$$\begin{aligned} u_{k+1} &= q_{k+1}(G)u_0 \\ &= \frac{2T_k(\xi)}{T_{k+1}(\xi)} l(G) q_k(G) u_0 - \frac{T_{k-1}(\xi)}{T_{k+1}(\xi)} q_{k-1}(G) u_0 \\ &= \frac{2T_k(\xi)}{T_{k+1}(\xi)} l(G) u_k - \frac{T_{k-1}(\xi)}{T_{k+1}(\xi)} u_{k-1}. \end{aligned} \tag{5.13}$$

将$u_i = \tilde{x}_i - x_*$代入 (5.13)，并注意到 $x_* = Gx_* + c$，即有

$$\tilde{x}_{k+1} = \frac{2T_k(\xi)}{T_{k+1}(\xi)} l(G)\tilde{x}_k - \frac{T_{k-1}(\xi)}{T_{k+1}(\xi)}\tilde{x}_{k-1} + \frac{4}{\beta-\alpha}\frac{T_k(\xi)}{T_{k+1}(\xi)}c. \tag{5.14}$$

令

$$\rho_{k+1} = 2\xi\frac{T_k(\xi)}{T_{k+1}(\xi)}, \quad \nu = \frac{2}{2-\beta-\alpha}.$$

则(5.14)可写作

$$\tilde{x}_{k+1} = \rho_{k+1}[\nu(G\tilde{x}_k + c) + (1-\nu)\tilde{x}_k] + (1-\rho_{k+1})\tilde{x}_{k-1}. \tag{5.15}$$

此外，易证

$$\rho_{k+1}-\frac{1}{4\xi^2}\rho_{k+1}\rho_k=1.$$

因此，我们可以按如下方式递推地计算ρ_k：

$$\rho_1=2,$$

$$\rho_{k+1}=\left(1-\frac{1}{4\xi^2}\rho_k\right)^{-1}. \tag{5.16}$$

这样，对给定的初值x_0，就可由公式（5.15）和(5.16)产生我们所需的序列$\{\tilde{x}_k\}_{k=0}^{\infty}$。这一迭代过程可总结如下：

（1）给定x_0，令$\rho_1=2$；

（2）计算：

$$\begin{aligned}x_1&=Gx_0+c,\\ \xi&=(2-\beta-\alpha)/(\beta-\alpha),\\ \nu&=2/(2-\beta-\alpha);\end{aligned} \tag{5.17}$$

（3）迭代：

$$\rho_{k+1}=\left(1-\frac{1}{4\xi^2}\rho_k\right)^{-1},$$

$$x_{k+1}=\rho_{k+1}[\nu(Gx_k+c)+(1-\nu)x_k]+(1-\rho_{k+1})x_{k-1}$$
$$(k=1,2,\cdots).$$

这一迭代法称作 **Chebyshev 半迭代法**，它是一个二步不定常的线性迭代法。

在实用中，常常取定一个正整数m，采用(5.17)迭代m次产生x_m之后，再以x_m作为初值x_0重新迭代m次，这样周而复始地进行下去直到满足要求为止。这种迭代法称为**循环 Chebyshev 半迭代法**. 从(5.8) 容易看出，循环Chebyshev 半迭代法收敛与否，最终归结为矩阵$q_m(G)$的谱半径是否小于1.

注意到

$$\max_{\alpha<t<\beta}\left|T_m\left(\frac{2t-\alpha-\beta}{\beta-\alpha}\right)\right|=1$$

和

$$\xi=\frac{2-\beta-\alpha}{\beta-\alpha}>1,$$

并利用 $t>1$时，

$$T_m(t)=[(t-\sqrt{t^2-1})^m+(t+\sqrt{t^2-1})^m]/2,$$

就可得

$$\rho(q_m(G))=\max_{\mu\in\lambda(G)}|q_m(\mu)|\leqslant\max_{\alpha\leqslant t\leqslant\beta}|q_m(t)|$$

$$=(T_m(\xi))^{-1}=\frac{2\sigma^m}{1+\sigma^{2m}}<1,$$

其中 $\sigma=\xi-\sqrt{\xi^2-1}=(\xi+\sqrt{\xi^2-1})^{-1}<1$。

这表明循环 Chebyshev 半迭代法在 G 的特征值均为小于 1 的实数时总是收敛的。

2．设 G 的特征值都是实数，但有的特征值大于 1．此时，原迭代法一定是不收敛的．由于迭代法是相容的，故 $I-G$ 非奇异，从而 1 必然不是 G 的特征值．因此，必存在 $\alpha<\beta<1$，使得

$$\lambda(G)\subset[\alpha,\beta]\cup[2-\beta,2-\alpha].$$

令 $G_1=G(2I-G)$，则

$$\lambda(G_1)\subset[\alpha(2-\alpha),\beta(2-\beta)]\subset(-\infty,1).$$

又

$$I-G_1=(I-G)^2,$$

故我们容易构造出迭代矩阵为 G_1 的相容的迭代法．于是对 G_1应用第一种情况所得到的迭代法，即可产生收敛的循环 Chebyshev 半迭代法。

3．设 G 的特征值不全是实数，但它包含在由如下椭圆界定的区域 $\mathscr{D}$ 之内：

$$\frac{(x-\xi)^2}{\alpha^2}+\frac{y^2}{\beta^2}=1,$$

其中 ξ,α,β 是实数，且满足 $\alpha>\beta>0$和$\alpha+\xi<1$，如图5.1所示.

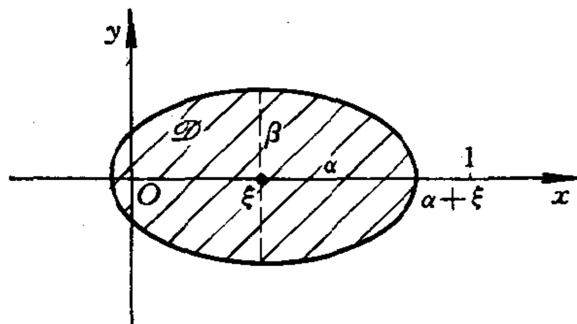


图 5.1

此时，我们考虑如下的优化问题：求 $q_k\in\mathscr{P}_k^{(1)}$使得

$$\max_{z\in\mathscr{D}}|q_k(z)|=\min_{p_k\in\mathscr{P}_k^{(1)}}\max_{z\in\mathscr{D}}|p_k(z)|.$$

此问题亦有唯一的解：

$$q_k(z)=\frac{T_k((z-\xi)/\gamma)}{T_k((1-\xi)/\gamma)},$$

其中 T_k仍是 k 次 Chebyshev 多项式，$\gamma=(\alpha^2-\beta^2)^{1/2}$. 利用Chebyshev 多项式的三项递推公式易得

$$q_0(z)\equiv 1,$$

$$q_1(z)\equiv\nu z-\nu+1,$$

$$q_{k+1}(z)=\rho_{k+1}(\nu z+1-\nu)q_k(z)+(1-\rho_{k+1})q_{k-1}(z)\qquad(k=1,2,\cdots),$$

其中

$$\nu=(1-\xi)^{-1},\qquad\rho_{k+1}=\frac{2\zeta T_k(\zeta)}{T_{k+1}(\zeta)},\quad\zeta=\frac{1-\xi}{\gamma}.$$

此时，Chebyshev 半迭代就变成

$$z_k=Gx_k-x_k+c,$$

$$x_{k+1}=\rho_{k+1}(\nu z_k+x_k)+(1-\rho_{k+1})x_{k-1}.$$

详细推导过程读者可模仿第一种情况进行，亦可参阅文献[1]和[49].

将 Chebyshev 多项式加速技巧与具体的迭代法相结合就可得

到相应的 Chebyshev 半迭代法。例如与前面的四种古典迭代法联合使用就得到了所谓的 J-SI,GS-SI,SOR-SI和 SSOR-SI 迭代法。由于只要 A 是对称的，就有 SSOR 的迭代矩阵 $\mathscr{S}_\omega$的特征值皆为实数，因此现在经常使用的是 SSOR-SI 迭代法。

Chebyshev 半迭代法的主要困难在于如何去高效地估计所需的参数（如 α,β,γ,ξ 等）。对某些应用领域中的实际问题，近来已有不少文章探求有关参数的估计方法。由于篇幅所限，这里不再赘述，有兴趣的读者可参阅文献[38]。

习　题

1．对于矩阵

$$A=\begin{bmatrix}4 & -1 & 0\\ -1 & 4 & -1\\ 0 & -1 & 4\end{bmatrix},$$

给出对应的点迭代矩阵 J，$\mathscr{L}_1,\mathscr{L}_\omega$和$\mathscr{S}_\omega$，并确定对应的 SOR和SSOR迭代的最优参数 ω，然后再比较它们的谱半径的大小。

2．如果存在非奇异矩阵W使得 $W(I-G)W^{-1}$是对称正定矩阵，则称迭代法(1.2)是可对称化的。现假定迭代法(1.2)是可对称化的。试证：

(1) G 的特征值皆为小于 1 的实数；

(2) 存在γ，使得

$$x_{k+1}=\gamma(Gx_k+d)+(1-\gamma)x_k,$$

是一个收敛的迭代法(这一迭代法称作(1.2)的一种外推方法)。

3．证明：对于SSOR 迭代法有

$$\rho(\mathscr{S}_\omega)\geqslant|1-\omega|^2,$$

并由此给出 SSOR 迭代法收敛的必要条件。

4．如果 $A=M-N$ 满足 $M^{-1}\geqslant 0$，$N\geqslant 0$，则称其为A的一个正则分裂。若 $A^{-1}\geqslant 0$，且 $A=M-N$是A的一个正则分裂，试证：

$$\rho(M^{-1}N)=\frac{\rho(A^{-1}N)}{1+\rho(A^{-1}N)}.$$

5. 设A为如下的块三对角阵

$$A=\begin{bmatrix} D_1 & C_2 & & 0 \\ B_2 & D_2 & \ddots & \\ & \ddots & \ddots & C_s \\ 0 & & B_s & D_s \end{bmatrix},$$

其中$D_i\in\mathbb{R}^{n_i\times n_i}$非奇异，$n_1+\cdots+n_s=n$。试证：对任意的$\mu\in\mathbb{C}\backslash\{0\}$，有

$$\det\left(D-\mu C_L-\frac{1}{\mu}C_U\right)=\det(D-C_L-C_U),$$

其中

$$D=\operatorname{diag}(D_1,\cdots,D_s),$$

$$C_L=-\begin{bmatrix} 0 & & & 0 \\ B_2 & 0 & & \\ & \ddots & \ddots & \\ 0 & & B_s & 0 \end{bmatrix},\qquad C_U=-\begin{bmatrix} 0 & C_2 & & 0 \\ & 0 & \ddots & \\ & & \ddots & C_s \\ 0 & & & 0 \end{bmatrix}.$$

6. 对于第5题所述的矩阵A，证明：当$\omega\neq 1$时，$\lambda\in\lambda(\mathscr{L}_\omega)$的充分必要条件是存在$\mu\in\lambda(J)$使得

$$\lambda=\frac{1}{4}[\omega\mu+(\omega^2\mu^2-4\omega+4)^{1/2}]^2.$$

7. 设形如第5题所述的矩阵A使对应的Jacobi 迭代阵J的特征值均为实数；并假定$\rho(J)<1$。试证：

(1) $R(\mathscr{L}_1)=2R(J)$;

(2) $\rho(\mathscr{L}_{\omega_b})=\inf\limits_{0<\omega<2}\rho(\mathscr{L}_\omega)=\omega_b-1$，其中

$$\omega_b=\frac{2}{1+\sqrt{1-\rho(J)^2}};$$

(3) $2\rho(J)(R(\mathscr{L}_1))^{1/2}\leqslant R(\mathscr{L}_{\omega_b})\leqslant R(\mathscr{L}_1)+2(R(\mathscr{L}_1))^{1/2}$,

这里假定 $R(\mathscr{L}_1)\leqslant 3$。

8. 设 $A=[a_{ij}]\in\mathbb{R}^{n\times n}$是非负矩阵。证明:

$$\min_i\sum_{j=1}^{n}a_{ij}\leqslant\rho(A)\leqslant\max_i\sum_{j=1}^{n}a_{ij}.$$

9. 设 $A=[a_{ij}]\in\mathbb{R}^{n\times n}$是严格对角占优的。试证:

$$|\det(A)|\geqslant\prod_{i=1}^{n}\left(|a_{ii}|-\sum_{j\neq i}|a_{ij}|\right).$$

第五章　共轭梯度法

§1　最速下降法

考虑线性方程组

$$Ax = b \tag{1.1}$$

的求解问题，其中 A 是给定的 n 阶对称正定矩阵，b 是给定的 n 维向量，x 是待求的 n 维向量。为此，我们定义二次泛函

$$\varphi(x) = \frac{1}{2}x^{\mathrm{T}}Ax - b^{\mathrm{T}}x. \tag{1.2}$$

容易验证：φ 的梯度为

$$\varphi'(x) = Ax - b, \tag{1.3}$$

φ 的Hessian 矩阵为 A。因此，φ 有唯一的极小点 $x_* = A^{-1}b$，从而有

$$Ax_* = b \Longleftrightarrow \varphi(x_*) = \min_x \varphi(x). \tag{1.4}$$

这表明求解方程组(1.1）与求二次泛函 φ 的极小点是等价的。早期，人们总是先把求 φ 的极小点的问题化为求解方程组（1.1)的问题，然后通过解(1.1）来求得 φ 的极小点。而这里，我们将把这个过程反过来，即把求解方程组(1.1）的问题化为求泛函 φ 的极小点的问题，然后通过求 φ 的极小点而求得方程组（1.1）的解。

求 φ 的极小点的最简单而有效的方法就是**最速下降法**，即从某点 x_0出发，逐步产生一串点 $x_0, x_1, \cdots, x_m, \cdots$，使 $\varphi(x_0) > \varphi(x_1) > \cdots > \varphi(x_m) > \cdots$以“最快的速度”下降到 φ 的极小值。具体的做法是，在求得 x_k之后，x_{k+1}是沿着 φ 在 x_k的最速下降方向，即负梯度方向 $r_k = -(Ax_k - b)$，求 φ 的最小值而得到的，即 $x_{k+1} =$

$x_k+\alpha_k\gamma_k$满足

$$\varphi(x_{k+1})=\min_{\alpha}\varphi(x_k+\alpha r_k). \tag{1.5}$$

因此，确定 x_{k+1}的过程也就是求解一个一元函数的极小点的过程，这很容易用初等微分学的方法来解决。

现令

$$f(\alpha)=\varphi(x_k+\alpha r_k),\qquad \alpha\in\mathbb{R}.$$

则易知

$$f'(\alpha)=r_k^{\mathrm{T}}(-r_k+\alpha Ar_k). \tag{1.6}$$

由 $f'(\alpha)=0$可求得 f 的唯一的极小点是

$$\alpha_{k+1}=r_k^{\mathrm{T}}r_k/r_k^{\mathrm{T}}Ar_k. \tag{1.7}$$

这样，我们就得到了最速下降法的迭代公式：

$$\begin{aligned}&\alpha_k=r_{k-1}^{\mathrm{T}}r_{k-1}/r_{k-1}^{\mathrm{T}}Ar_{k-1},\\&x_k=x_{k-1}+\alpha_k r_{k-1},\\&r_k=b-Ax_k\\&\qquad(k=1,2,\cdots).\end{aligned} \tag{1.8}$$

对于这一迭代法有如下的收敛性定理。

定理1.1 设 A 的特征值为$0<\lambda_1\leqslant\cdots\leqslant\lambda_n$。则由迭代法(1.8)产生的点列$\{x_k\}_{k=0}$满足

$$\|x_k-x_*\|_A\leqslant\left(\frac{\lambda_n-\lambda_1}{\lambda_n+\lambda_1}\right)^k\|x_0-x_*\|_A, \tag{1.9}$$

其中 $x_*=A^{-1}b$，$\|\cdot\|_A$按第一章的(3.3)式定义。

为了给出这一定理的证明，我们先证一个引理，这一引理在今后经常用到。

引理1.1 设 A 的特征值为 $0<\lambda_1\leqslant\cdots\leqslant\lambda_n$，$p(t)$是 t 的一个多项式。则

$$\|p(A)x\|_A\leqslant\max_{1\leqslant i\leqslant n}|p(\lambda_i)|\,\|x\|_A,\qquad\forall\, x\in\mathbb{R}^n.$$

证明 设 $y_1,y_2,\cdots,y_n$是 A 的对应于 $\lambda_1,\lambda_2,\cdots,\lambda_n$的特征向量

所构成的 $\mathbb{R}^n$ 的一组标准正交基. 则对任意的 $x \in \mathbb{R}^n$ 有 $x = \sum_{i=1}^{n} \beta_i y_i$, 从而有

$$
\begin{aligned}
x^{\mathrm{T}} p(A) A p(A) x &= \left(\sum_{i=1}^{n} \beta_i p(\lambda_i) y_i\right)^{\mathrm{T}} \left(\sum_{i=1}^{n} \beta_i \lambda_i p(\lambda_i) y_i\right) \\
&= \sum_{i=1}^{n} \lambda_i \beta_i^2 p^2(\lambda_i) \\
&\leqslant \max_{1 \leqslant i \leqslant n} p^2(\lambda_i) \sum_{i=1}^{n} \lambda_i \beta_i^2 \\
&= \max_{1 \leqslant i \leqslant n} p^2(\lambda_i) x^{\mathrm{T}} A x.
\end{aligned}
$$

于是

$$\|p(A)x\|_A \leqslant \max_{1 \leqslant i \leqslant n} |p(\lambda_i)| \, \|x\|_A.$$

定理1.1的证明 由 x_k 满足

$$\varphi(x_k) \leqslant \varphi(x_{k-1} + \alpha r_{k-1}), \quad \forall \alpha \in \mathbb{R}$$

可得

$$
\begin{aligned}
&(x_k - x_*)^{\mathrm{T}} A (x_k - x_*) \\
&\quad \leqslant (x_{k-1} + \alpha r_{k-1} - x_*)^{\mathrm{T}} A (x_{k-1} + \alpha r_{k-1} - x_*) \\
&\quad = [(I - \alpha A)(x_{k-1} - x_*)]^{\mathrm{T}} A [(I - \alpha A)(x_{k-1} - x_*)], \\
&\qquad\qquad \forall \alpha \in \mathbb{R}.
\end{aligned}
\tag{1.10}
$$

记 $p_\alpha(t) = 1 - \alpha t$, 并应用引理1.1, 从 (1.10) 可得

$$
\begin{aligned}
\|x_k - x_*\|_A &\leqslant \|p_\alpha(A)(x_{k-1} - x_*)\|_A \\
&\leqslant \max_{1 \leqslant i \leqslant n} |p_\alpha(\lambda_i)| \, \|x_{k-1} - x_*\|_A
\end{aligned}
\tag{1.11}
$$

对一切的 $\alpha \in \mathbb{R}$ 成立, 再利用 Chebyshev 多项式的性质, 可得

$$\min_{\alpha} \max_{\lambda_1 \leqslant t \leqslant \lambda_n} |1 - \alpha t| = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}. \tag{1.12}$$

将(1.12)代入(1.11)即得

$$\|x_k - x_*\|_A \leqslant \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} \|x_{k-1} - x_*\|_A. \tag{1.13}$$

注意到(1.13) 对于一切的自然数 k 都成立，立即知(1.9)成立。

定理 1.1 表明，从任一初始向量 x_0出发，由最速下降法产生的点列$\{x_k\}_{k=0}^{\infty}$总是收敛到方程组(1.1)的解，其收敛的快慢由量$\frac{\lambda_n-\lambda_1}{\lambda_n+\lambda_1}$的大小来决定。

综合上面的讨论，我们就得到了求解方程组 (1.1)的如下算法。

算法1.1

(1) 输入 A,b,x_0和精度 ε。

(2) $r:=b-Ax_0$, $x:=x_0$。

(3) $\alpha:=r^{\mathrm{T}}r/r^{\mathrm{T}}Ar$,

$x:=x+\alpha r$,

$r:=b-Ax$。

(4) 如果$\|r\|_\infty<\varepsilon$，则输出 x，结束；否则转步 (3)。

这一算法有简单易用，可充分利用 A 的稀疏性等优点，但当$\lambda_1\ll\lambda_n$时收敛速度变得非常之慢，以致于完全不适用。如何来加快这一算法的收敛速度，是我们今后几节将要讨论的问题。

此外，在这一章今后几节里，如果没有特别说明，我们所讨论的矩阵 A 都是正定的，所谈到的 φ 均指由(1.2) 所定义的二次泛函。

§2 二次泛函的几何性质

为了进一步考察和分析最速下降法的特性，寻找改进的途径，这一节我们对二次泛函 φ 的性质作进一步的研究。为此，先引进几个几何概念。

设$p,q\in\mathbb{R}^n$。我们说 p 和 q 是**互相共轭的**(或更确切地讲，是互相 A 共轭的)，是指它们满足 $p^{\mathrm{T}}Aq=0$。

设 $p_1,\cdots,p_{n-k}$ 是 $\mathbb{R}^n$ 中 $n-k$ 个线性无关的向量，$\delta_1,\cdots,\delta_{n-k}$ 是给定的 $n-k$ 个实数。我们称满足

$$p_i^{\mathrm{T}}x=\delta_i,\quad i=1,2,\cdots,n-k \tag{2.1}$$

的 x 的全体所构成的 $\mathbb{R}^n$ 中的点集为一个 k **维超平面**，记作

$$\pi_k:\ p_i^{\mathrm{T}}x=\delta_i,\qquad i=1,2,\cdots,n-k, \tag{2.2}$$

其中的 p_i 称作 π_k 的**法向量**。

从定义容易看出，一个 k 维超平面实质上是 $n-k$ 个法向量线性无关的通常意义下的超平面的交集。

此外，由 $p_1,\cdots,p_{n-k}$ 线性无关的假定知，齐次方程组

$$p_i^{\mathrm{T}}x=0,\qquad i=1,2,\cdots,n-k$$

有且仅有 k 个线性无关的非零解。现任取它的一组线性无关的非零解 $u_1,u_2,\cdots,u_k$ 和 π_k 中的一点 x_0，则易证 $x\in\pi_k$ 的充分必要条件是存在实数 $\alpha_1,\alpha_2,\cdots,\alpha_k$，使得

$$x=x_0+\alpha_1u_1+\cdots+\alpha_ku_k.$$

令 $U=[u_1,\cdots,u_k]$，则 π_k 亦可表为

$$\pi_k:\ x=x_0+Uy,\qquad y\in\mathbb{R}^k. \tag{2.3}$$

此时，我们亦称 π_k 是经过点 x_0 由 $u_1,\cdots,u_k$ 张成的 k 维超平面。(2.3)也说明一个 k 维超平面就是一个 k 维子空间的平移。

我们说一个向量 p 包含在一个 k 维超平面 π_k 内，是指这一向量的起点和终点都在 π_k 上，这等价于 p 有一点在 π_k 上，而且满足

$$p_i^{\mathrm{T}}p=0,\qquad i=1,2,\cdots,n-k, \tag{2.4}$$

其中 $p_1,\cdots,p_{n-k}$ 是 π_k 的法向量；我们说一个向量 q 垂直于 π_k，是指 q 与 π_k 内的任一向量都正交，这等价于 q 满足

$$q^{\mathrm{T}}u_i=0,\qquad i=1,2,\cdots,k, \tag{2.5}$$

其中 $u_1,\cdots,u_k$ 是张成 π_k 的 k 个向量。

设 π_k 是任一以 $p_1,\cdots,p_{n-k}$ 为法向的 k 维超平面，则称由 $A^{-1}p_1,\cdots,A^{-1}p_{n-k}$所张成的超平面

$$\hat{\pi}_{n-k}:\ x=x_*+\alpha_1A^{-1}p_1+\cdots+\alpha_{n-k}A^{-1}p_{n-k} \qquad (2.6)$$

为π_k的**共轭超平面**，其中$x_*=A^{-1}b$是方程组(1.1)的解。

这里需指出的是，上面所定义的共轭超平面这一概念与方程组(1.1)有关。

此外，从定义易知：π_k 内的向量与 $\hat{\pi}_{n-k}$ 内的向量是互相共轭的，而且若π_k是由$u_1,\cdots,u_k$张成的，则 $Au_1,\cdots,Au_k$就是 $\hat{\pi}_{n-k}$ 的法向。

设 B 是 m阶对称正定矩阵，γ是任一正数，x_0是 $\mathbb{R}^m$中的任一点。我们称由满足

$$(x-x_0)^{\mathrm{T}}B(x-x_0)=\gamma \qquad (2.7)$$

的 x 的全体所构成的$\mathbb{R}^m$中的点集为 **$m-1$维超椭球面**；x_0 称作这个超椭球的**中心**。记作

$$E_{m-1}:\ (x-x_0)^{\mathrm{T}}B(x-x_0)=\gamma.$$

由上述定义，易证点集

$$E_{n-1}:\ \varphi(x)=\gamma \quad (\gamma>\varphi(x_*)) \qquad (2.8)$$

是一个中心为 $x^*=A^{-1}b$ 的 $n-1$ 维超椭球面。

现在我们来考虑 φ 在任一给定的超平面上的极小点的特征。

定理2.1 设

$$\pi_k:\ x=x_0+Uy, \qquad y\in\mathbb{R}^k$$

是任一给定的 k 维超平面，其中 $U\in\mathbb{R}_k^{n\times k}$。则

(1) φ 在 π_k 上的极小点是唯一的；

(2) φ 在 π_k 上的极小点正好是 π_k 截 $n-1$维超椭球面(2.8)所得到的 $k-1$ 维超椭球面

$$E_{k-1}=E_{n-1}\cap\pi_k \qquad (2.9)$$

所界定的超椭球的中心；

(3) $\tilde{x}$ 是 φ 在 π_k 上的极小点的充分必要条件是 φ 在 $\tilde{x}$ 的梯度 $\varphi'(\tilde{x})$ 垂直于 π_k。

证明 φ 在 π_k 上的限制为二次泛函

$$\psi(y) = \varphi(x_0 + Uy) = \frac{1}{2}y^{\mathrm{T}}By - g^{\mathrm{T}}y + \varphi(x_0), \qquad y \in \mathbb{R}^k,$$

其中 $B = U^{\mathrm{T}}AU$，$g = -U^{\mathrm{T}}\varphi'(x_0) = U^{\mathrm{T}}r_0$，$r_0 = b - Ax_0$。由 A 正定和 U 满秩知 B 亦是正定的。因此，ψ 在 $\mathbb{R}^k$ 内有唯一的极小点 $\tilde{y} = B^{-1}g$，从而 φ 在 π_k 上有唯一的极小点 $\tilde{x} = x_0 + U\tilde{y}$。再注意到 $\tilde{y}$ 是 $k-1$ 维超椭球面

$$\widetilde{E}_{k-1}:\ \psi(y) = \delta \quad (\delta \geqslant \psi(\tilde{y}))$$

所界定的超椭球的中心，即知 $\tilde{x} = x_0 + U\tilde{y}$ 是 $k-1$ 维超椭球面

$$E_k = E_{n-1} \cap \pi_k = \{x:\ x = x_0 + Uy,\ y \in \widetilde{E}_k\}$$

所界定的超椭球的中心。

此外，注意到

$$\psi'(y) = By - g = U^{\mathrm{T}}(AUy - r_0) = U^{\mathrm{T}}\varphi'(x_0 + Uy), \qquad (2.10)$$

即知定理的(3)成立。证毕。

定理2.2 设 $p_1, \cdots, p_{n-k}$ 是给定的 $n-k$ 个线性无关的向量，则 φ 在任一以 $p_1, \cdots, p_{n-k}$ 为法向的 k 维超平面 π_k 上的极小点必位于 π_k 的共轭超平面 $\hat{\pi}_{n-k}$ 上。

证明 设 $\tilde{x}$ 是 φ 在 π_k 上的极小点。则由定理2.1知，$\varphi'(\tilde{x}) = A\tilde{x} - b$ 垂直于 π_k。而 π_k 的法向量是 $p_1, \cdots, p_{n-k}$，故有

$$A\tilde{x} - b \in \mathrm{span}\{p_1, \cdots, p_{n-k}\},$$

即存在常数 $\alpha_1, \cdots, \alpha_{n-k}$，使

$$A\tilde{x} - b = \alpha_1 p_1 + \alpha_2 p_2 + \cdots + \alpha_{n-k} p_{n-k}.$$

于是有

$$\tilde{x} = A^{-1}b + \alpha_1 A^{-1}p_1 + \alpha_2 A^{-1}p_2 + \cdots + \alpha_{n-k}A^{-1}p_{n-k},$$

即 $\tilde{x}\in\pi_{n-k}$. 证毕.

现在我们来从几何上解释，当A的最大特征值λ_n远远大于它的最小特征值λ_1时，最速下降法的收敛速度为什么有时会变得非常之慢.

从几何上来看，求φ的极小点就是寻求$n-1$维超椭球面E_{n-1}（如(2.8)所定义）所界定的椭球的中心；而最速下降法的每一步，就相当于从一个给定的点$\tilde{x}$出发，沿着超椭球面

$$\tilde{E}_{n-1}: \ \varphi(x)=\varphi(\tilde{x})$$

的内法线方向$\tilde{r}=b-A\tilde{x}$，寻找φ的极小值，即寻找φ在一维超平面

$$\pi_1: \ x=\tilde{x}+\alpha\tilde{r}, \qquad \alpha\in\mathbb{R}$$

上的极小点；定理2.1表明，这也就是寻找π_1在椭球面$\tilde{E}_{n-1}$内之线段的中点$\hat{x}$；当A的最大特征值λ_n远远大于它的最小特征λ_1时，椭球面$\tilde{E}_{n-1}$变得非常扁平；如果$\tilde{x}$位于它的较平坦的一面时，其内法线方向$\tilde{r}$就与$\tilde{x}$和椭球中心$x_*=A^{-1}b$的连线方向几乎是垂直的，这样$\hat{x}$与x_*之间的距离就差不多等于$\tilde{x}$与x_*之间的距离，从而使得最速下降法的收敛速度变得非常之慢. $n=2$时的示意图如图2.1所示.

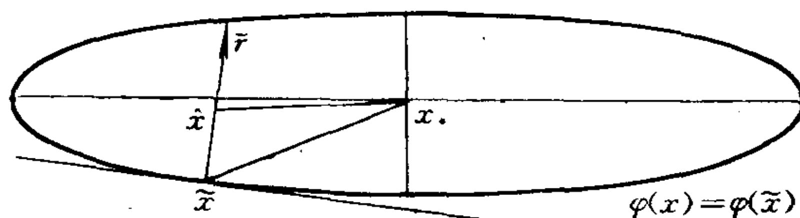


图 2.1

§3 共轭梯度法及其基本性质

上一节对最速下降法的几何解释说明，就我们所解决的问题而言，负梯度方向并非最合适的方向，有时它与目标的偏差太

大。那么，能否选出比负梯度方向更好而计算量又不是很大的方向作为新的下降方向呢？答案是肯定的，下面将要介绍的共轭梯度法，就是其中的一种。

首先，随便选取一点 $x_0\in\mathbb{R}^n$ 作为初值，并计算 $r_0=b-Ax_0$。第一步，仍取 r_0 作为下降方向，记作 $p_0=r_0$。现假定由此出发已进行了 k 步（$k\geq 0$），求得了 x_k 和下降方向 p_k。下一步，我们先在直线（即一维超平面）

$$l:\quad x=x_k+\alpha p_k,\qquad \alpha\in\mathbb{R}$$

上求 φ 的极小点 x_{k+1}。完全与最速下降法类似，可得

$$x_{k+1}=x_k+\alpha_{k+1}p_k,\tag{3.1}$$

其中

$$\alpha_{k+1}=r_k^{\mathrm{T}}p_k/p_k^{\mathrm{T}}Ap_k,\tag{3.2}$$
$$r_k=b-Ax_k.$$

然后，再来分析如何选取一个新的下降方向 p_{k+1}。由于我们希望 p_{k+1} 比 φ 在点 x_{k+1} 的负梯度方向 $r_{k+1}=b-Ax_{k+1}$ 更好，且确定 p_{k+1} 的运算量又不很大，因此我们自然应在经过点 x_{k+1} 且包含 r_{k+1} 的二维超平面内考虑 p_{k+1} 的选取问题。这样的二维超平面应该如何选取呢？如果 $r_{k+1}=0$，则 x_{k+1} 就是方程组(1.1)的解，当然迭代就此结束；如果 $r_{k+1}\neq 0$，则 r_{k+1} 和 p_k 是互相垂直的两个非零向量（直接验证有 $r_{k+1}^{\mathrm{T}}p_k=0$）。这就给我们以启示，所要选择的二维超平面的一个简单而自然的选法应该是选取经过 x_{k+1} 由 r_{k+1} 和 p_k 张成的二维超平面

$$\pi_2:\quad x=x_{k+1}+\alpha r_{k+1}+\beta p_k$$

在 π_2 内的最佳方向，当然应该是 φ 在这一平面内的极小点 $\tilde{x}$ 和 x_{k+1} 的连线方向。从定理2.1知，$\tilde{x}$ 就是 π_2 与超椭球面

$$E_{n-1}:\quad \varphi(x)=\varphi(x_{k+1})$$

相交所得到的椭圆 S_2 的中心（参见图3.1）；而且定理2.1还告诉

我们，$\tilde{x}$ 应满足

$$r_{k+1}^{\mathrm{T}}\varphi'(\tilde{x})=p_k^{\mathrm{T}}\varphi'(\tilde{x})=0. \tag{3.3}$$

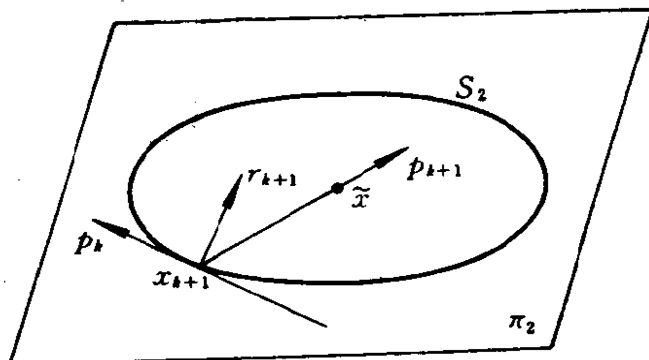


图 3.1

而 $\tilde{x}\in\pi_2$，故有

$$\tilde{x}=x_{k+1}+\alpha_0 r_{k+1}+\beta_0 p_k,$$

从而有

$$\begin{aligned}\varphi'(\tilde{x})&=A\tilde{x}-b\\&=Ax_{k+1}+\alpha_0 Ar_{k+1}+\beta_0 Ap_k-b\\&=-r_{k+1}+\alpha_0 Ar_{k+1}+\beta_0 Ap_k.\end{aligned} \tag{3.4}$$

将(3.4)代入(3.3)，并整理得

$$\begin{cases}\alpha_0 r_{k+1}^{\mathrm{T}}Ar_{k+1}+\beta_0 r_{k+1}^{\mathrm{T}}Ap_k=r_{k+1}^{\mathrm{T}}r_{k+1}, & (3.5)\\ \alpha_0 p_k^{\mathrm{T}}Ar_{k+1}+\beta_0 p_k^{\mathrm{T}}Ap_k=0. & (3.6)\end{cases}$$

易知，这一方程组有且仅有唯一的一组解(α_0,β_0)，并可断言$\alpha_0\neq 0$。这是因为，若 $\alpha_0=0$，则由(3.6) 知 $\beta_0=0$，从而由 (3.5) 知 $r_{k+1}^{\mathrm{T}}r_{k+1}=0$，这与 $r_{k+1}\neq 0$ 的假定相矛盾。因此，我们就可选取新的下降方向为

$$p_{k+1}=\frac{1}{\alpha_0}(\tilde{x}-x_{k+1})=r_{k+1}+\frac{\beta_0}{\alpha_0}p_k. \tag{3.7}$$

令 $\beta_{k+1}=\beta_0/\alpha_0$，则由(3.6) 知

$$\beta_{k+1} = -p_k^T A r_{k+1} / p_k^T A p_k. \tag{3.8}$$

这样，我们就以不太大的运算量确定了一个新的方向 p_{k+1}，而且几何直观表明，这一方向上 φ 的最小值要比在 r_{k+1} 上的最小值来的小，也就是说这样选定的新方向比用负梯度方向要好一些。

综述上面的讨论，就得到了如下的迭代公式：

$$\begin{aligned} \alpha_k &= r_{k-1}^T p_{k-1} / p_{k-1}^T A p_{k-1}, \\ x_k &= x_{k-1} + \alpha_k p_{k-1}, \\ r_k &= r_{k-1} - \alpha_k A p_{k-1}, \\ \beta_k &= -p_{k-1}^T A r_k / p_{k-1}^T A p_{k-1}, \\ p_k &= r_k + \beta_k p_{k-1} \end{aligned} \tag{3.9}$$

$$(k = 1, 2, \cdots),$$

其中 $x_0 \in \mathbb{R}^n$ 任意给定，$p_0 = r_0 = b - Ax_0$。

迭代法(3.9) 的最基本性质可归纳为如下定理。

定理3.1 设对任给的初值 x_0 已用迭代公式(3.9)迭代了 m 次($m<n$)。则对任意的 k，$1 \leqslant k \leqslant m$，有

(1) $p_i^T r_k = 0$，$i = 0, 1, \cdots, k-1$； (3.10)

(2) $r_i^T r_k = 0$，$i = 0, 1, \cdots, k-1$； (3.11)

(3) $p_i^T A p_k = 0$，$i = 0, 1, \cdots, k-1$； (3.12)

(4) $\operatorname{span}\{r_0, r_1, \cdots, r_k\} = \operatorname{span}\{p_0, p_1, \cdots, p_k\}$
$= \operatorname{span}\{r_0, Ar_0, \cdots, A^k r_0\}$。 (3.13)

证明 对 k 应用数学归纳法.

当 $k=1$ 时，直接验证知，$p_0^T r_1 = 0$，$r_0^T r_1 = 0$，$p_0^T A p_0 = 0$，以及

$$\operatorname{span}\{r_0, r_1\} = \operatorname{span}\{p_0, p_1\} = \operatorname{span}\{r_0, Ar_0\}$$

成立。

现假定对 $k=j$，$1 \leqslant j < m$ 已证定理成立，下面考虑 $k=j+1$ 的情形。

(1) 利用等式 $r_{j+1} = r_j - \alpha_{j+1} A p_j$ 及归纳法假定，有

$$p_i^T r_{j+1} = p_i^T r_j - \alpha_{j+1} p_i^T A p_j = 0, \qquad i = 0,1,\cdots,j-1.$$

又由于

$$p_j^T r_{j+1} = p_j^T r_j - \frac{r_j^T p_j}{p_j^T A p_j} p_j^T A p_j = 0,$$

故对 $k=j+1$ 亦有 (3.10) 成立。

(2) 利用归纳法假定有

$$\mathrm{span}\{r_0,r_1,\cdots,r_j\} = \mathrm{span}\{p_0,p_1,\cdots,p_j\},$$

而由(1)所证知，r_{j+1} 垂直于子空间 $\mathrm{span}\{p_0,\cdots,p_j\}$，从而有 r_{j+1} 垂直于子空间 $\mathrm{span}\{r_0,\cdots,r_j\}$。因此 (3.11) 亦对 $k=j+1$ 成立。

(3) 利用 $p_{j+1} = r_{j+1} + \beta_{j+1} p_j$，有

$$p_i^T A p_{j+1} = p_i^T A r_{j+1} + \beta_{j+1} p_i^T A p_j. \tag{3.14}$$

而由 $r_{i+1} = r_i - \alpha_{i+1} A p_i$，有

$$A p_i = \frac{1}{\alpha_{i+1}} (r_i - r_{i+1}). \tag{3.15}$$

将 (3.15) 代入 (3.14)，并利用归纳假定和 (2) 所证结论，即有

$$\begin{aligned} p_i^T A p_{j+1} &= \frac{1}{\alpha_{i+1}} r_{j+1}^T (r_i - r_{i+1}) + \beta_{j+1} p_i^T A p_j \\ &= 0, \qquad i = 0,1,\cdots,j-1. \end{aligned}$$

而

$$p_j^T A p_{j+1} = 0$$

是显然的，从而(3.12)亦对 $k=j+1$ 成立。

(4) 由归纳法假定知

$$r_j, p_j \in \mathrm{span}\{r_0, Ar_0, \cdots, A^j r_0\},$$

因此，有

$$r_{j+1} = r_j - \alpha_{j+1} A p_j \in \mathrm{span}\{r_0, Ar_0, \cdots, A^j r_0, A^{j+1} r_0\};$$

从而，有

$$p_{j+1}=r_{j+1}+\beta_{j+1}p_j\in \text{span}\{r_0,Ar_0,\cdots,A^{j+1}r_0\}.$$

而(2)和(3)所证的结论表明，向量组 $r_0,\cdots,r_{j+1}$ 和 $p_0,\cdots,p_{j+1}$ 都是线性无关的，因此有(3.13)亦对 $k=j+1$ 成立。

由归纳法原理即知定理对一切 $k(1\leqslant k\leqslant m)$ 成立。证毕。

通常记

$$\kappa(A,r_0,j)=\text{span}\{r_0,Ar_0,\cdots,A^{j-1}r_0\},$$

并称之为 **Krylov 子空间**。

定理3.1表明：按（3.9）迭代所产生的剩余向量 $r_0,r_1,\cdots,r_m$ 是互相正交的；方向向量 $p_0,p_1,\cdots,p_m$ 是互相共轭的；它们分别是 Krylov 子空间 $\kappa(A,r_0,m+1)$的一组正交基和一组共轭正交基。因此，最多迭代 $m=n-1$步，就必有 $r_m=0$，即在有限步之内必可求得(1.1) 的精确解。

此外，从定理3.1的(1)和(2)立即知，

$$r_k^{\mathrm T}r_k=r_k^{\mathrm T}(r_{k-1}-\alpha_kAp_{k-1})=-\alpha_kr_k^{\mathrm T}Ap_{k-1},\tag{3.16}$$

$$\begin{aligned}r_{k-1}^{\mathrm T}r_{k-1}&=r_{k-1}^{\mathrm T}(p_{k-1}-\beta_{k-1}p_{k-2})=r_{k-1}^{\mathrm T}p_{k-1}\\&=\alpha_kp_{k-1}^{\mathrm T}Ap_{k-1}.\end{aligned}\tag{3.17}$$

由(3.17)得

$$\alpha_k=r_{k-1}^{\mathrm T}r_{k-1}/p_{k-1}^{\mathrm T}Ap_{k-1}.\tag{3.18}$$

由(3.16)，(3.17)和(3.8)，得

$$\beta_k=r_k^{\mathrm T}r_k/r_{k-1}^{\mathrm T}r_{k-1}.\tag{3.19}$$

用(3.18)和(3.19)分别代替(3.9)中 α_k 和 β_k 的计算公式，就得到如下算法。

算法3.1

(1) 输入 A,b 和x_0；$r_0:=b-Ax_0$。

(2) 如果 $r_0=0$，则输出 x_0，结束；否则

$$p_0:=r_0,\qquad \rho_0:=r_0^{\mathrm T}r_0,\qquad k:=1.$$

(3) $\alpha_k:=\rho_{k-1}/p_{k-1}^{\mathrm T}Ap_{k-1}$,

$x_k:=x_{k-1}+\alpha p_{k-1}$,

$r_k := r_{k-1} - \alpha_k A p_{k-1}$,

$\rho_k := r_k^T r_k$,

$\beta_k := \rho_k / \rho_{k-1}$,

$p_k := r_k + \beta_k p_{k-1}$。

(4) 如果 $r_k = 0$，则输出 x_k，结束；否则 $k := k+1$，转步(3)。

这一算法称作共轭梯度法，简称 CG（Conjugate Gradient）法。这样命名的缘由可从下面的定理明白。

定理3.2 设算法3.1执行到 $k = m$ 时结束。则对任意的 $1 \leqslant k < m$，有

(1) x_k 是 φ 在 k 维超平面

$$\pi_k: \quad x = x_0 + \xi_0 p_0 + \xi_1 p_1 + \cdots + \xi_{k-1} p_{k-1}$$

上的极小点；

(2) p_k 与 φ 在点 x_k 的负梯度方向 $r_k = b - Ax_k$ 在 π_k 的共轭超平面 $\hat{\pi}_{n-k}$ 上的正交投影同向。

证明 由定理2.1知，欲证(1)成立，只需证 $\varphi'(x_k) = -r_k$ 垂直于超平面 π_k 即可，而这只需证

$$r_k^T p_i = 0, \qquad i = 0, 1, \cdots, k-1, \tag{3.20}$$

即可。(3.20)正是定理3.1的(1)所述的结论，从而(1)得证。

由于 $Ap_0, \cdots, Ap_{k-1}$ 是 $\hat{\pi}_{n-k}$ 的法向，$x_k \in \hat{\pi}_{n-k}$（定理2.2），所以 $\hat{\pi}_{n-k}$ 可表示为

$$\hat{\pi}_{n-k}: \quad x = x_k + Vy, \quad y \in \mathbb{R}^{n-k},$$

其中 $V = [v_1, \cdots, v_{n-k}] \in \mathbb{R}^{n \times (k-k)}$ 满足 $V^T V = I$，且

$$p_i^T A v_j = 0, \qquad i = 0, 1, \cdots, k-1, j = 1, \cdots, n-k。 \tag{3.21}$$

现在考虑 φ 在 $\hat{\pi}_{n-k}$ 上的限制

$$\psi(y) = \varphi(x_k + Vy), \qquad y \in \mathbb{R}^{n-k}.$$

令 $r = -\psi'(0)$，即 r 为 ψ 在 $y = 0$ 的最速下降方向，并定义

$$p = Vr, \tag{3.22}$$

那么利用 $\psi'(0)=V^{\mathrm{T}}\varphi'(x_k)$，可得

$$p=VV^{\mathrm{T}}r_k, \tag{3.23}$$

其中 $r_k=-\varphi'(x_k)=b-Ax_k$ 是 φ 在 x_k 的最速下降方向。

从(3.23)可知向量 p 在超平面 $\hat{\pi}_{n-k}$ 内，且 $q=r_k-p$ 与平面 $\hat{\pi}_{n-k}$ 垂直，即

$$q^{\mathrm{T}}v_i=0,\quad i=1,2,\cdots,n-k. \tag{3.24}$$

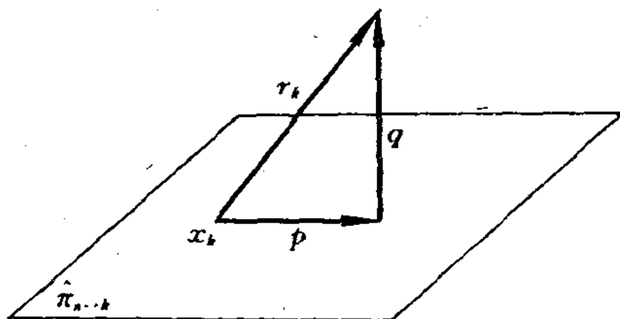


图 3.2

这也就是说 p 是 r_k 在平面 $\hat{\pi}_{n-k}$ 内的正交投影（见图3.2）。因此，只要证明 p_k 与 p 同向，即证存在正数 α 使

$$p_k=\alpha p, \tag{3.25}$$

则定理3.2的（2）即得证。

由(3.24)和(3.21)知，

$$q\in\operatorname{span}\{Ap_0,\cdots,Ap_{k-1}\},$$

即存在常数 $\tilde{\beta}_0,\cdots,\tilde{\beta}_{k-1}$，使

$$r_k-p=\tilde{\beta}_0Ap_0+\cdots+\tilde{\beta}_{k-1}Ap_{k-1}. \tag{3.26}$$

注意到

$$Ap_i=\frac{1}{\alpha_{i+1}}(r_i-r_{i+1}),\qquad i=0,1,\cdots,k-1,$$

$$r_i=p_i-\beta_ip_{i-1},\qquad i=0,1,\cdots,k,$$

从(3.26)即知，存在常数 $\tilde{\alpha}_0,\cdots,\tilde{\alpha}_k$，使得

$$p=\tilde{\alpha}_0p_0+\tilde{\alpha}_1p_1+\cdots+\tilde{\alpha}_kp_k.$$

上式两边左乘 $p_j^{\mathrm{T}}A$，注意到 p 正交于 $\hat{\pi}_{n-k}$ 的法向 $Ap_0,\cdots,Ap_{k-1}$ 和 p 与 $p_0,\cdots,p_{k-1},p_k$ 互相共轭，就有

$$0=\tilde{\alpha}_jp_j^{\mathrm{T}}Ap_j,\qquad j=0,1,\cdots,k-1.$$

又 $p_j^{\mathrm{T}}Ap_j\neq0$，故必有

$$\tilde{\alpha}_j = 0, \qquad j = 0,1,\cdots,k-1.$$

于是

$$p = \tilde{\alpha}_k p_k. \tag{3.27}$$

为证 $\tilde{\alpha}_k > 0$，我们考虑 φ 在 x_k 沿方向 p 的方向导数 $\varphi'_p(x_k)$。由 p 的定义和方向导数的定义易知

$$\varphi'_p(x_k) = \psi'_r(0) = -[\psi'(0)]^{\mathrm{T}}[\psi'(0)] < 0, \tag{3.28}$$

其中 $\psi'_r(0)$表示 ψ 在 $y=0$ 沿方向 r 的方向导数，另一方面，又有

$$\varphi'_p(x_k) = p^{\mathrm{T}}\varphi'(x_k) = -\tilde{\alpha}_k p_k^{\mathrm{T}} r_k = -\tilde{\alpha}_k r_k^{\mathrm{T}} r_k. \tag{3.29}$$

结合（3.28）和(3.29)即知 $\tilde{\alpha}_k > 0$，从而（3.25）得证。证毕。

定理3.2表明，共轭梯度法的第 $k+1$ 步实质上是沿着 φ 在 π_k 的共轭超平面 $\tilde{\pi}_{n-k}$ 内于点 x_k 处的负梯度方向 p 求 φ 的极小点 x_{k+1}。由此不难明白算法3.1称作共轭梯度法的缘故了。

§4 实用共轭梯度法及其收敛性

4.1 实用共轭梯度法

上一节导出的共轭梯度法，在理论上，由于所得到的剩余向量 r_k 的相互正交性，而推出至多迭代 $n-1$ 步就能得到方程组(1.1)的精确解；但在实际使用时，由于误差的出现，使得 r_k 之间的正交性很快损失，以致于其有限步终止性已不再成立。因此，在实际上，我们只能将共轭梯度法作为一种迭代法使用。这样，就得到如下流行的共轭梯度法。

算法4.1

（1）输入 A, b 和 x_0；

$$x := x_0,\quad r := Ax_0,\quad \rho_0 := r^{\mathrm{T}} r,\quad k := 1.$$

（2）如果 $k=1$，则 $p := r$；否则

$$\beta := \rho_{k-1}/\rho_{k-2}, \qquad p := r + \beta p.$$

(3) $w:=Ap$, $\alpha:=\rho_{k-1}/p^{\mathrm{T}}w$,

$x:=x+\alpha p$, $r:=r-\alpha w$,

$\rho_k:=r^{\mathrm{T}}r$.

(4) 如果 $\rho_k<\rho_0\varepsilon$，则输出 x，结束；否则 $k:=k+1$，转步(2)。

这一算法，每迭代一次，只需作一次矩阵与向量乘法运算和 $5n$ 次数量乘法运算；在存储上仅需再增加$4n$个存储单元存放向量 r,x,p 和 w 即可。此外，在实际使用时，上述算法中 ρ_k的下标亦是不必要的。

共轭梯度法作为一种实用的迭代法，它主要有下面的优点：

(1) 可充分利用 A 的稀疏性，存储时只需存储 A 的非零元素即可；

(2) 不需预先估计别的参数就可以计算，这一点不像 Chebyshev 半迭法和松弛法等；

(3) 每次迭代所需的计算，主要是向量之间的运算，便于并行化。

此外，还需指出的一点是，系数矩阵为 Hilbert 矩阵的线性方程组是典型的病态方程组，无论是直接法还是迭代法，一般都很难得到较为精确的结果，而数值试验发现共轭梯度法却能得到较为精确的解。详细情况可参阅黄友谦（参见文献[5]）主编的《数值试验》第七章。这样，是否可认为共轭梯度法是对付病态问题的有效方法呢？还有待于对各种不同类型的问题进行大量的数值试验，才能作出结论。

4.2 收敛性分析

将共轭梯度法作为一种迭代法，它的收敛性怎样呢？这是本节下面将要讨论的主要问题。

记

$$u_k=x_k-x_*, \tag{4.1}$$

其中 x_k 是算法4.1第 k 次迭代产生的极小化向量，$x_* = A^{-1}b$；u_k 称之为算法4.1第 k 步产生的**误差向量**。

引理4.1 算法4.1第 k 步产生的误差向量满足：

$$\|u_k\|_A \leqslant \min_{q_k \in \mathscr{P}_k^{(0)}} \max_{\lambda \in \lambda(A)} |q_k(\lambda)| \|u_0\|_A, \tag{4.2}$$

其中 $\mathscr{P}_k^{(0)}$ 表示满足 $q_k(0)=1$的次数不超过 k 的实系数多项式的全体。

证明 由定理3.1和3.2知，x_k 是二次泛函 φ 在 k 维超平面

$$\pi_k: \ x = x_0 + \xi_0 p_0 + \cdots + \xi_{k-1} p_{k-1}$$

上的极小点，且

$$\text{span}\{p_0, \cdots, p_{k-1}\} = \text{span}\{r_0, Ar_0, \cdots, A^{k-1}r_0\}.$$

由此不难推出

$$\varphi(x_k) = \min_{q_k \in \mathscr{P}_k^{(0)}} \varphi(x_* + q_k(A)u_0). \tag{4.3}$$

由 φ 的定义容易验证(4.3) 等价于

$$\|u_k\|_A = \min_{q_k \in \mathscr{P}_k^{(0)}} \|q_k(A)u_0\|_A. \tag{4.4}$$

利用引理1.1,有

$$\|q_k(A)u_0\|_A \leqslant \max_{\lambda \in \lambda(A)} |q_k(\lambda)| \|u_0\|_A, \tag{4.5}$$

对任意的 $q_k \in \mathscr{P}_k^{(0)}$成立。将（4.5）代入（4.4）即得（4.2）。证毕。

（4.2）表明，对于任意的 $q_k \in \mathscr{P}_k^{(0)}$，量 $\max\limits_{\lambda \in \lambda(A)} |q_k(\lambda)|$都可作为相对误差

$$\|u_k\|_A / \|u_0\|_A \tag{4.6}$$

的上界。在这众多的上界中，要选取一个最佳上界，就需求解下面的优化问题：求 $q_k^* \in \mathscr{P}_k^{(0)}$，使

$$\max_{\lambda \in \lambda(A)} |q_k^*(\lambda)| = \min_{q_k \in \mathscr{P}_k^{(0)}} \max_{\lambda \in \lambda(A)} |q_k(\lambda)|. \tag{4.7}$$

如第四章第五节曾讲过的求解这样一个离散型的优化问题是相当

困难的。因此，人们通常考虑与（4.7）相关的一个连续型问题：求 $q_k^* \in \mathscr{P}_k^{(0)}$，使

$$\max_{\lambda\in[a,b]}|q_k^*(\lambda)| = \min_{q_k\in\mathscr{P}_k^{(0)}}\ \max_{\lambda\in[a,b]}|q_k(\lambda)|, \tag{4.8}$$

其中 $a=\lambda_{\min}(A)$，$b=\lambda_{\max}(A)$。由著名的 Chebyshev 多项式逼近定理知，（4.8）的解是

$$q_k^*(\lambda;a,b)=T_k\left(\frac{b+a-2\lambda}{b-a}\right)\Big/T_k\left(\frac{b+a}{b-a}\right), \tag{4.9}$$

其中 $T_k(x)$ 是 k 次 Chebyshev 多项式。

现令

$$F(a,b;k)=\max_{\lambda\in[a,b]}|q_k^*(\lambda;a,b)|,$$

则有

$$F(a,b;k)=\left(T_k\left(\frac{b+a}{b-a}\right)\right)^{-1}=\frac{2\sigma^k}{1+\sigma^{2k}}, \tag{4.10}$$

其中

$$\sigma=(\sqrt{b}-\sqrt{a})/(\sqrt{b}+\sqrt{a}).$$

再注意到

$$\frac{\sigma^k}{1+\sigma^{2k}}=\frac{(\sqrt{b}-\sqrt{a})^k(\sqrt{b}+\sqrt{a})^k}{(\sqrt{b}+\sqrt{a})^{2k}+(\sqrt{b}-\sqrt{a})^{2k}}$$

$$\leqslant\frac{(\sqrt{b}-\sqrt{a})^k}{(\sqrt{b}+\sqrt{a})^k}$$

$$=\left(\sqrt{\frac{b}{a}}-1\right)^k\Big/\left(\sqrt{\frac{b}{a}}+1\right)^k,$$

就可得如下关于共轭梯度法的收敛性定理。

定理4.1 在引理4.1的假设下，有

$$\|u_k\|_A\leqslant2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k\|u_0\|_A, \tag{4.11}$$

其中 $\kappa=\kappa_2(A)$表示A的谱条件数.

如果我们对A的谱分布作进一步的假定，(4.11)还可以改进. 这里我们只介绍其中的几个结果.

定理4.2 设u_k是算法4.1产生的误差向量. 则有:

(1) 如果 $\lambda(A)\subset\{\lambda_1,\cdots,\lambda_p\}\cup[a_1,b_1]$，其中 $\lambda_i\in\lambda(A)$，$\lambda_i<a_1$，那么

$$\|u_k\|_A\leqslant F(a_1,b_1;k-p)\prod_{i=1}^{p}\frac{b_1}{\lambda_i}\|u_0\|_A; \tag{4.12}$$

(2) 如果 $\lambda(A)\subset[a_2,b_2]\cup\{\lambda'_1,\cdots,\lambda'_p\}$，其中 $\lambda'_i\in\lambda(A)$，$\lambda'_i>b_2$，那么

$$\|u_k\|_A\leqslant F(a_2,b_2;k-p)\|u_0\|_A; \tag{4.13}$$

(3) 如果 $\lambda(A)\subset\{\lambda_1,\cdots,\lambda_p\}\cup[a_3,b_3]\cup\{\lambda'_1,\cdots,\lambda'_p\}$，其中 $\lambda_i,\lambda'_i\in\lambda(A)$，$\lambda'_i>b_3$，$\lambda_i<a_3$，那么

$$\|u_k\|_A\leqslant\frac{1}{4}\prod_{i=1}^{p}\frac{\lambda'_i}{\lambda_i}\left(1-\frac{\lambda_i}{\lambda'_i}\right)^2F(a_3,b_3;k-2)\|u_0\|_A, \tag{4.14}$$

其中 $F(a,b;k)$按(4.10)定义.

证明 (1) 取

$$q_k(\lambda)=\prod_{i=1}^{p}\left(1-\frac{\lambda}{\lambda_i}\right)q^*_{k-p}(\lambda;a_1,b_1),$$

其中 q^*_{k-p}按(4.9)定义，则易知 $q_k\in\mathscr{P}_k^{(0)}$. 再由

$$q_k(\lambda_j)=0,\qquad j=1,2,\cdots,p$$

和

$$\left|1-\frac{\lambda}{\lambda_i}\right|<\frac{b_1}{\lambda_i},\qquad a_1\leqslant\lambda\leqslant b_1,$$

即有

$$\max_{\lambda\in\lambda(A)}|q_k(\lambda)|\leqslant\max_{\lambda\in[a_1,b_1]}|q_k(\lambda)|\leqslant\prod_{i=1}^{p}\frac{b_1}{\lambda_i}F(a_1,b_1;k-p).$$

再利用引理4.1即有(4.12)成立.

(2) 取

$$\tilde{q}_k(\lambda)=\prod_{i=1}^{p}\left(1-\frac{\lambda}{\lambda'_i}\right)q^*_{k-p}(\lambda;a_2,b_2),$$

则 $\tilde{q}_k\in\mathscr{P}_k^{(0)}$，且有

$$\max_{\lambda\in\lambda(A)}|\tilde{q}_k(\lambda)|\leqslant\max_{\lambda\in[a_2,b_2]}|\tilde{q}(\lambda)|\leqslant F(a_2,b_2;k-p).$$

再应用引理4.1即知（4.13)成立。

(3) 取

$$\hat{q}_k(\lambda)=\prod_{i=1}^{p}\frac{(\lambda_i-\lambda)(\lambda'_i-\lambda)}{\lambda_i\lambda'_i}\cdot q^*_{k-2p}(\lambda;a_3,b_3),$$

则有 $q_k\in\mathscr{P}_k^{(0)}$，而且

$$\max_{\lambda\in\lambda(A)}|\hat{q}_k(\lambda)|\leqslant\max_{\lambda\in[a_3,b_3]}|\hat{q}(\lambda)|$$

$$\leqslant\frac{1}{4}\prod_{i=1}^{p}\frac{\lambda'_i}{\lambda_i}\left(1-\frac{\lambda_i}{\lambda'_i}\right)^2F(a_3,b_3;k-2p).$$

再用引理4.1又有（4.14)成立。证毕。

定理4.1和4.2表明，共轭梯度法收敛的快慢依赖于系数矩阵的谱分布情况；当系数矩阵的条件数很小，或其谱大部分集中在一点附近而仅有少数几个远离此点时，可以期望用算法 4.1 迭代很少几步就会得到高精度的近似解。大量的数值试验表明，在这些情况下，往往需要比理论上估计的迭代次数更少的迭代次数就可得到所需精度的近似解。这一现象就是所谓的共轭梯度法的“超线性收敛性”。

§5　预优共轭梯度法

上一节对共轭梯度法的收敛性分析表明，当系数矩阵的特征值较均匀地分布在一个很长的区间上时，共轭梯度法的收敛速度可能会变得很慢。大量的数值试验也充分证实了这一点。可是，

在实际应用中却经常遇到这种情况。因此，如何提高共轭梯度法的收敛速度就显得非常重要，并成为近二十多年来研究的一个重要课题。

定理4.1和4.2启发我们，如果能够选取一个非奇异矩阵 C，使 $\tilde{A}=C^{-1}AC^{-T}$ 的特征值分布在一个较小的区间内，或分布较为“集中”的话，那么应用共轭梯度法于方程组

$$\tilde{A}\tilde{x}=\tilde{b}, \tag{5.1}$$

其中 $\tilde{x}=C^{T}x$，$\tilde{b}=C^{-1}b$，将会有较快的收敛速度，进而可以提高求解(1.1) 的速度。

将共轭梯度法应用于方程组 (5.1)的迭代公式具体地写出来就是：

$$\begin{aligned}
&\alpha_k=\tilde{r}_{k-1}^{T}\tilde{r}_{k-1}/\tilde{p}_k\tilde{A}\tilde{p}_k,\\
&\tilde{x}_k=\tilde{x}_{k-1}+\alpha_k\tilde{p}_k,\\
&\tilde{r}_k=\tilde{r}_{k-1}-\alpha_k\tilde{A}\tilde{p}_k, \qquad (k=1,2,\cdots),\\
&\beta_k=\tilde{r}_k^{T}\tilde{r}_k/\tilde{r}_{k-1}^{T}\tilde{r}_{k-1},\\
&\tilde{p}_{k+1}=\tilde{r}_k+\beta_k\tilde{p}_k
\end{aligned}$$

其中 $\tilde{x}_0$是任取的初值，$\tilde{r}_0=\tilde{b}-\tilde{A}\tilde{x}_0$，$\tilde{p}_1=\tilde{r}_0$。

按照上述公式直接迭代，需计算 $\tilde{A}=C^{-1}AC^{-T}$和 $\tilde{b}$，而且还需将迭代得到的近似解 $\tilde{x}_k$.通过变换$x_k=C^{-T}\tilde{x}_k$变成方程组(1.1)的近似解。实际上这些都是不必要的，作变换

$$x_k=C^{-T}\tilde{x}_k, \quad r_k=c\tilde{r}_k, \quad p_k=C^{-T}\tilde{p}_k,$$

并记 $M=CC^{T}$，代入上面的各式，即可得到如下的算法。

算法5.1

(1) 输入 A,M,b 和 x_0；

$r_0:=b-Ax_0$，$z_0:=M^{-1}r_0$，$p_1:=z_0$，

$\rho_0:=r_0^{T}z_0$，$k:=1$。

(2) $w:=Ap_k$，$\alpha_k:=\rho_{k-1}/p_k^{T}w$，

$x_k:=x_{k-1}+\alpha_kp_k$，$r_k:=r_{k-1}-\alpha_kw$，

$$z_k := M^{-1}r_k, \quad \rho_k := r_k^{\mathrm{T}} z_k,$$
$$\beta_k := \rho_k / \rho_{k-1}, \quad p_{k+1} := z_k + \beta_k p_k.$$

(3) 如果 $\rho_k < \rho_0 \varepsilon$,则输出 x_k,结束;否则 $k := k+1$ 转步(2).

这一算法称作**预优共轭梯度法**,也有的作者称作**预条件共轭梯度法**,简称 PCG(Preconditioned Conjugate Gradient)法;M称作**预优矩阵**.

容易从共轭梯度法的基本性质推知,这样得到的方向向量p_k和剩余向量r_k满足:

$$r_i^{\mathrm{T}} M^{-1} r_j = 0, \quad i \neq j; \tag{5.2}$$
$$p_i^{\mathrm{T}} A p_j = 0, \quad i \neq j. \tag{5.3}$$

此外,易知算法 5.1产生的 x_k满足

$$\|x_k - x_*\|_A \leqslant 2\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right) \|x_0 - x_*\|_A, \tag{5.4}$$

其中 κ 为$M^{-1}A$ 的最大特征值与最小特征值之比,$x_* = A^{-1}b$.

现在的问题是M应该怎样选取最好.从算法5.1可以看出,迭代的每一步都要解一个形如 $Mz = r$ 的方程组.因此,一个较好的预优矩阵M应该具有如下的特征:

(1) M对称正定;

(2) M应该与A的稀疏性差不多;

(3) $M^{-1}A$ (即 $\tilde{A} = C^{-1}AC^{-\mathrm{T}}$) 的特征值分布"集中";

(4) 形如 $Mz = r$ 的方程组容易求解,即M应具有某些特殊形状,如块对角,或三角矩阵的乘积等.

要找到同时具备这样一些特性的预优矩阵M并非一件易事.下面我们从几个侧面来提供一些选择M的有用信息.

为此,我们来考虑算法5.1的一种等价形式.由

$$x_{k-1} = x_{k-2} + \alpha_{k-1} p_{k-1},$$

可得

$$p_{k-1}=\frac{1}{\alpha_{k-1}}(x_{k-1}-x_{k-2}).\tag{5.5}$$

于是有

$$p_k=z_{k-1}+\beta_{k-1}p_{k-1}=z_{k-1}+\frac{\beta_{k-1}}{\alpha_{k-1}}(x_{k-1}-x_{k-2}),\tag{5.6}$$

从而有

$$x_k=x_{k-1}+\alpha_k p_k=x_{k-1}+\alpha_k\left[z_{k-1}+\frac{\beta_{k-1}}{\alpha_{k-1}}(x_{k-1}-x_{k-2})\right].\tag{5.7}$$

对(5.7) 稍加整理，可写成

$$x_k=x_{k-2}+\omega_k(\gamma_{k-1}z_{k-1}+x_{k-1}-x_{k-2}),\tag{5.8}$$

其中 ω_k和 γ_k是参数。下面来推导这两个参数的计算公式。(5.8) 两边左乘 $-A$，并加上 b，然后再利用 $Mz_i=r_i=b-Ax_i$，即有

$$Mz_k=Mz_{k-2}-\omega_k(\gamma_{k-1}Az_{k-1}+Mz_{k-2}-Mz_{k-1}).\tag{5.9}$$

再注意到

$$r_i^{\mathrm{T}}M^{-1}r_j=0,\ \Rightarrow z_i^{\mathrm{T}}Mz_j=0,\ i\neq j,$$

在(5.9) 两边分别左乘 z_{k-1}^{T}和 z_{k-2}^{T}，得

$$0=-\omega_k(\gamma_{k-1}z_{k-1}^{\mathrm{T}}Az_{k-1}-z_{k-1}^{\mathrm{T}}Mz_{k-1}),\tag{5.10}$$

$$0=z_{k-2}^{\mathrm{T}}Mz_{k-2}-\omega_k(\gamma_{k-1}z_{k-2}^{\mathrm{T}}Az_{k-1}+z_{k-2}^{\mathrm{T}}Mz_{k-2}).\tag{5.11}$$

从(5.10)得

$$\gamma_{k-1}=z_{k-1}^{\mathrm{T}}Mz_{k-1}/z_{k-1}^{\mathrm{T}}Az_{k-1}.\tag{5.12}$$

从(5.11)得

$$\omega_k=(1+\gamma_{k-1}z_{k-2}^{\mathrm{T}}Az_{k-1}/z_{k-2}^{\mathrm{T}}Mz_{k-2})^{-1}.\tag{5.13}$$

为了尽可能减少计算 ω_k的运算量，我们给出 $z_{k-2}^{\mathrm{T}}Az_{k-1}$的另一种等价表述。在(5.9) 中将下标 k 换作$k-1$，并且在两边左乘 z_{k-1}^{T}，即有

$$z_{k-1}^{\mathrm{T}}Mz_{k-1}=-\omega_{k-1}\gamma_{k-2}z_{k-1}^{\mathrm{T}}Az_{k-2}.\tag{5.14}$$

由(5.14)和(5.13)可得

$$\omega_k = \left(1 - \frac{\gamma_{k-1}}{\gamma_{k-2}} \cdot \frac{1}{\omega_{k-1}} \cdot \frac{z_{k-1}^{\mathrm{T}} M z_{k-1}}{z_{k-2}^{\mathrm{T}} M z_{k-2}}\right)^{-1}. \qquad (5.15)$$

这样，我们就得到了理论上与算法5.1等价的另一种算法。

算法5.2

(1) 输入 A, b, M 和 x_0；

$r_0 := b - Ax_0$，$z_0 := M^{-1} r_0$，$\rho_0 := z_0^{\mathrm{T}} r_0$，

$\gamma_0 := \rho_0 / z_0^{\mathrm{T}} A z_0$。

(2) $x_1 := x_0 + \gamma_0 z_0$，$r_1 := b - Ax_1$，$z_1 := M^{-1} r_1$，

$\omega_1 := 1$，$\rho_1 := z_1^{\mathrm{T}} r_1$，$k := 2$。

(3) $w := Az_{k-1}$，$\gamma_{k-1} := \rho_{k-1} / z_{k-1}^{\mathrm{T}} w$，

$$\omega_k := \left(1 - \frac{\gamma_{k-1}}{\gamma_{k-2}} \cdot \frac{1}{\omega_{k-1}} \cdot \frac{\rho_{k-1}}{\rho_{k-2}}\right)^{-1},$$

$x_k := x_{k-2} + \omega_k (\gamma_{k-1} z_{k-1} + x_{k-1} - x_{k-2})$，

$r_k := b - Ax_k$，$z_k := M^{-1} r_k$，

$\rho_k := z_k^{\mathrm{T}} r_k$。

(4) 如果 $\rho_k < \rho_0 \varepsilon$，则输出 x_k，结束；否则 $k := k+1$，转步(3)。

下面我们来看，将算法5.1改写成算法5.2，在寻找M的问题上，会给我们提供一些什么样的信息。

在基本迭代公式(5.8)中，令 $\omega_k = \gamma_{k-1} = 1$，则有

$$x_k = z_{k-1} + x_{k-1}. \qquad (5.16)$$

再将 $z_{k-1} = M^{-1}(b - Ax_{k-1})$ 代入(5.16)，并左乘M，得

$$Mx_k = Nx_{k-1} + b, \qquad (5.17)$$

其中

$$N = M - A,$$

即

$$A = M - N. \qquad (5.18)$$

这样，从另一角度来讲，算法 5.2 可以看作是对应于基本迭

代法（5.17）的一种加速方法．这启发我们可充分利用现有的迭代法的分裂方式来产生M．例如，基于SSOR迭代法的分裂方式，可取预优矩阵M为

$$M=(D-\omega C_L)D^{-1}(D-\omega C_U),$$

其中$A=D-C_L-C_U$，D是块对角矩阵，且对称正定易于求逆，ω是松弛因子．这样选取M得到的PCG算法通常称作SSOR-CG法，它在实用上是一种很有效的方法，详细的讨论可参见文献[38]．

再如，许多椭圆偏微分方程的离散化形式是一个系数矩阵为

$$A=\begin{bmatrix} M_1 & F \\ F^{\mathrm{T}} & M_2 \end{bmatrix}\begin{matrix} m \\ m \end{matrix}$$
$$\qquad\; m \quad\;\; m$$

的线性方程组，其中A对称正定，且形如$M_1z_1=r_1$和$M_2z_2=r_2$的方程组容易求解．此时，自然取

$$M=\begin{bmatrix} M_1 & 0 \\ 0 & M_2 \end{bmatrix}$$

为预优矩阵，从而导出了非常有效的块Jacobi迭代法的加速方法——J-CG法．

在结束本节之前，我们需顺便指出的是，算法5.2的另一个优点是可以较容易地推广到一类非对称的线性方程组的求解问题．因此，亦称这一算法为**广义共轭梯度法**．

设线性方程组

$$Ax=b$$

的系数矩阵A非奇异，且有如下分裂

$$A=M-N,$$

其中M是对称正定矩阵，N为反对称矩阵．则容易看出，此时，可以完全照搬算法5.2，而得到求解这类方程组的算法．

对任意的系数矩阵A，恒有

$$A=\frac{1}{2}(A+A^{\mathrm{T}})-\frac{1}{2}(A^{\mathrm{T}}-A)。$$

令$M=\frac{1}{2}(A+A^{\mathrm{T}})$，$N=\frac{1}{2}(A^{\mathrm{T}}-A)$，显然有$M$对称，$N$反对称。因此，只要$\frac{1}{2}(A+A^{\mathrm{T}})$正定，则原则上就可以应用算法5.2。但这种情况下，M^{-1}未必是容易求出的。因而实际应用时仍有一定的困难，这是一个仍需进一步研究的课题。

§6 不完全分解预优技巧

大家知道，预优共轭梯度法成败的关键在于预优矩阵M选择的是否恰当。在共轭梯度法发展的初期，由于人们只是利用一些矩阵的简单分裂方式来选取预优矩阵M，以致于使得很多情况下收敛速度的提高并不十分明显，因而致使共轭梯度法自问世之后，二十多年来一直没有得到广泛的应用，直到七十年代后期，两位荷兰数学家 Meijerink和Van der Vorst 才将 Cholesky 分解法和共轭梯度巧妙地结合起来，提出了一种称作不完全分解预优共轭梯度法。这类方法，由于其运算量往往可与直接法媲美，而又具有迭代法节省内存的优点，因此受到人们的推崇，成为近十年来国内外研究的热点。这使得线性方程组的求解方法经历了相当长的平稳发展时期以后，终于发生了突破性的进展。

所谓不完全 Cholesky 分解，就是将A分解成

$$A=LL^{\mathrm{T}}+R, \tag{6.1}$$

其中L是下三角矩阵，R称作**剩余矩阵**。由于这里有R可以变化，因此L中哪些元素为零可以预先规定。这样，我们就可以要求L与A保持同样的稀疏性，或者具有我们希望的某种稀疏性，从而克服了完全 Cholesky 分解破坏A的稀疏性的缺点。

不完全分解预优共轭梯度法，就是先对A进行形如(6.1)的

不完全 Cholesky 分解；然后，以 $M = LL^T$ 作为预优矩阵来应用 PCG 方法．因此，为了使所得到的PCG方法收敛尽可能快，就需要LL^T尽可能地接近于 A．

形如(6.1) 的分解有相当大的灵活性．一是 L 的哪些元素为零可根据不同的要求来指定；二是 R 的选择上亦有相当大的自由度．近十几年来，人们针对不同类型的问题，利用分解 (6.1)的灵活性，已经提出了各种各样的不完全分解．这里，我们将着重介绍人们公认的并得到广泛应用的一种不完全分解——松弛不完全分解．

6.1 松弛不完全 LU 分解

设$A = [a_{ij}] \in \mathbb{R}^{n \times n}$是一给定的矩阵（不一定对称）．令

$$\mathscr{T}_n = \{(i,j):\ 1 \leqslant i \neq j \leqslant n\}, \tag{6.2a}$$

$$\mathscr{D}_A = \{(i,j) \in \mathscr{T}_n:\ a_{ij} \neq 0\}. \tag{6.2b}$$

对于$\mathscr{T}_n$的任一含有 $\mathscr{D}_A$的子集$\mathscr{T}$和任一参数 ω，$0 \leqslant \omega \leqslant 1$，我们说 A 有关于$\mathscr{T}$和 ω 的**松弛不完全LU分解**（简称RILU分解），是指矩阵 A 有如下分解

$$A = LU + R, \tag{6.3a}$$

其中 $L = [l_{ij}]$是单位下三角矩阵，且满足

$$l_{ij} = 0, \qquad (i,j) \notin \mathscr{T} \text{ 且 } i > j; \tag{6.3b}$$

$U = [u_{ij}]$是上三角矩阵，并满足

$$u_{ij} = 0, \qquad (i,j) \notin \mathscr{T} \text{ 且 } i < j; \tag{6.3c}$$

$R = [r_{ij}]$满足

$$r_{ij} = 0, \qquad (i,j) \in \mathscr{T}, \tag{6.3d}$$

$$r_{ii} = -\omega \sum_{j \neq i} r_{ij}, \qquad i = 1, 2, \cdots, n, \tag{6.3e}$$

其中的 ω 称作**松弛参数**．

在上述分解中，如果$\omega = 0$，则对应的分解就是著名的**不完全 LU 分解**(简称为 ILU)；如果 $\omega = 1$，则对应的分解就是通常人们

所讲的**修正不完全 LU 分解**（简称 MILU分解）。

对于给定的矩阵 A，如果它有松弛不完全 LU 分解，则我们可以用 Gauss 消去法稍加修正实现分解（6.3）。具体分解过程可叙述如下：记$A_1 = A$，对 $k = 1,2,\cdots,n-1$依次进行：

（1）Gauss 消去．用Gauss 变换将$A_k = [a_{ij}^{(k)}]$的第 k 列的第 k 个元素之下的元素全部消为零，即计算

$$\tilde{A}_k = L_k A_k = A_k - l_k a_k, \tag{6.4a}$$

其中

$$l_k = (0,\cdots,0,l_{k+1,k},\cdots,l_{nk})^{\mathrm{T}} \in \mathbb{R}^n, \tag{6.4b}$$

$$l_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)}, \qquad i = k+1,\cdots,n, \tag{6.4c}$$

$$a_k = e_k^T A_k = (0,\cdots,0,a_{kk}^{(k)},a_{k,k+1}^{(k)},\cdots,a_{kn}^{(k)}), \tag{6.4d}$$

$$L_k = I - l_k e_k^T \ (\text{Gauss 变换}); \tag{6.4e}$$

（2）修正．修正 $\tilde{A}_k = [\tilde{a}_{ij}^{(k)}]$，使其具有指定的稀疏性，并满足对角元素的特定要求，即计算

$$A_{k+1} = \tilde{A}_k - R_k, \tag{6.5a}$$

其中 $R_k = [r_{ij}^{(k)}]$定义作

$$r_{ij}^{(k)} = \begin{cases} \tilde{a}_{ij}^{(k)}, & k+1 \leqslant i \neq j \leqslant n \text{ 且 } (i,j) \notin \mathscr{T}, \\ -\omega \displaystyle\sum_{p \in \mathscr{P}(k,i)}^{n} \tilde{a}_{ip}^{(k)}, & k+1 \leqslant i = j \leqslant n, \\ 0, & \text{其他}, \end{cases} \tag{6.5b}$$

这里 $\mathscr{P}(k,i) = \{p:\ k+1 \leqslant p \leqslant n,\ p \neq i,\ (i,p) \notin \mathscr{T}\}$。

令

$$L = (L_{n-1}\cdots L_1)^{-1},\ U = A_n,\ R = \sum_{k=1}^{n-1} R_k. \tag{6.6}$$

则 L,U 和R 满足（6.3）的所有要求。事实上，由上述的消去修正过程，我们易知：

（1）$A_{k+1} = [a_{ij}^{(k+1)}]$具有如下形状

$$A_{k+1} = \begin{bmatrix} A_{11}^{(k+1)} & A_{12}^{(k+1)} \\ 0 & A_{22}^{(k+1)} \end{bmatrix}, \tag{6.7}$$

其中$A_{11}^{(k+1)}$是$k\times k$上三角阵，且当$(i,j)\in\mathscr{T}$且$i\neq j$时，$a_{ij}^{(k+1)}=0$，$k=0,1,\cdots,n-1$。

(2) U的第k行就是A_k的第k行，即

$$u_k=e_k^{\mathrm{T}}U=e_k^{\mathrm{T}}A_k=a_k,\qquad k=1,2,\cdots,n. \tag{6.8}$$

(3)
$$L=I+l_1e_1^{\mathrm{T}}+\cdots+l_{n-1}e_{n-1}^{\mathrm{T}}$$
$$=\begin{bmatrix}1&&&&\\l_{21}&1&&0&\\l_{31}&l_{32}&1&&\\\vdots&\vdots&\ddots&\ddots&\\l_{n1}&l_{n2}&\cdots&l_{n,n-1}&1\end{bmatrix}, \tag{6.9}$$

其中l_{ij}由（6.4c）给出。

因此，由（6.4)～(6.6）产生的L和U分别满足（6.3b）和（6.3c)。再将（6.4a）代入（6.5a)，即有

$$A_{k+1}=A_k-l_ka_k-R_k. \tag{6.10}$$

从而有

$$A_n=A_1-\sum_{k=1}^{n-1}l_ka_k-\sum_{k=1}^{n-1}R_k. \tag{6.11}$$

注意到$A_n=U$，$A_1=A$，以及

$$LU=U+l_1u_1+\cdots+l_{n-1}u_{n-1}$$
$$=U+l_1a_1+\cdots+l_{n-1}a_{n-1},$$

从(6.11)可知

$$A=LU+R. \tag{6.12}$$

此外，从R的定义（6.6)和(6.5b）知R满足（6.3d)和(6.3e)。

为了更清楚地了解这一分解过程，现在再举一例。

例6.1 设

$$A=\begin{bmatrix}3&-1&0&-2\\-2&4&-1&0\\0&0&1&-1\\-1&0&-1&2\end{bmatrix}.$$

求 A 关于$\mathscr{T}=\mathscr{D}_A$和 $\omega=1$的松弛不完全 LU 分解。

利用刚才介绍的消去修正法，要求的分解可用三步完成：

(1) $$\tilde{A}_1=L_1A=\begin{bmatrix}3 & -1 & 0 & -2\\ 0 & 10/3 & -1 & -4/3\\ 0 & 0 & 1 & -1\\ 0 & -1/3 & -1 & 4/3\end{bmatrix},$$

其中

$$L_1=I-l_1e_1^{\mathrm{T}},\qquad l_1=(0,-2/3,0,-1/3)^{\mathrm{T}},$$

$$A_2=\tilde{A}_1-R_1=\begin{bmatrix}3 & -1 & 0 & -2\\ 0 & 2 & -1 & 0\\ 0 & 0 & 1 & -1\\ 0 & 0 & -1 & 1\end{bmatrix},$$

其中

$$R_1=\begin{bmatrix}0 & 0 & 0 & 0\\ 0 & 4/3 & 0 & -4/3\\ 0 & 0 & 0 & 0\\ 0 & -1/3 & 0 & 1/3\end{bmatrix};$$

(2) $A_3=\tilde{A}_2=A_2$，$R_2=0$，$L_2=I$；

(3) $$\tilde{A}_3=L_3A_3=\begin{bmatrix}3 & -1 & 0 & -2\\ 0 & 2 & -1 & 0\\ 0 & 0 & 1 & -1\\ 0 & 0 & 0 & 0\end{bmatrix},$$

其中

$$L_3=I-l_3e_3^{\mathrm{T}},\qquad l_3=(0,0,0,-1)^{\mathrm{T}},$$

$$A_4=\tilde{A}_3,\quad R_3=0,$$

于是我们求得

$$L=\begin{bmatrix}1 & 0 & 0 & 0\\ -2/3 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ -1/3 & 0 & -1 & 1\end{bmatrix},\quad U=\begin{bmatrix}3 & -1 & 0 & -2\\ 0 & 2 & -1 & 0\\ 0 & 0 & 1 & -1\\ 0 & 0 & 0 & 0\end{bmatrix},$$

$$R=\begin{bmatrix}0 & 0 & 0 & 0\\ 0 & 4/3 & 0 & -4/3\\ 0 & 0 & 0 & 0\\ 0 & -1/3 & 0 & 1/3\end{bmatrix}.$$

直接验证可知 $A=LU+R$。

对于给定的矩阵 A 及 $\mathscr{T}$ 和 ω，容易看出，消去修正过程不中断的充分必要条件是

$$a_{kk}^{(k)}\neq 0,\qquad k=1,2,\cdots,n-1. \tag{6.13}$$

因此，自然要问：A 具有什么性质才能保证（6.13）成立呢？这里，我们只给出一个容易验证而且很多偏微分方程的离散化方程组都满足的条件。

定义6.1 设 $A=[a_{ij}]\in\mathbb{R}^{n\times n}$。如果它满足

$$|a_{ii}|\geqslant\sum_{j\neq i}|a_{ij}|,\qquad i=1,2,\cdots,n,$$

则称 A 是**对角占优矩阵**。

定义6.2 $A=[a_{ij}]\in\mathbb{R}^{n\times n}$ 称作是 $\hat{M}$ **矩阵**，如果它满足：

(1) $a_{ii}>0$，$i=1,2,\cdots,n-1$，$a_{nn}\geqslant 0$；

(2) $a_{ij}\leqslant 0$，$i\neq j$；

(3) $n(i)>i$，$i=1,2,\cdots,n-1$，

其中 $n(i)=\max\{j:\ 1\leqslant j\leqslant n,\ a_{ij}\neq 0\}$，即 $n(i)$ 是 A 的第 i 行元素中最后一个非零元素所在的列。

定理6.1 设 $A=[a_{ij}]\in\mathbb{R}^{n\times n}$ 是对角占优的 $\hat{M}$ 矩阵，则由消去修正过程（6.4）和（6.5）产生的矩阵 $A_1,A_2,\cdots,A_n$ 都是对角占优的 $\hat{M}$ 矩阵，而且有

(1) $a_{ij}^{(k+1)}\leqslant a_{ij}^{(k)}\leqslant 0$，$k+1\leqslant i\neq j\leqslant n$， (6.14a)

(2) $S_i^{(k+1)}\geqslant S_i^{(k)}\geqslant 0$，$i=k+1,\cdots,n$， (6.14b)

(3) $0<a_{ii}^{(k+1)}\leqslant a_{ii}^{(k)}$，$i=k+1,\cdots,n-1$， (6.14c)

(4) $0\leqslant a_{nn}^{(k+1)}\leqslant a_{nn}^{(k)}$， (6.14d)

其中 $S_i^{(k)}$ 表示 A_k 的第 i 行元素之和，即

$$S_i^{(k)}=\sum_{j=1}^{n}a_{ij}^{(k)}.$$

证明 用数学归纳法证之。

由$A_1=A$知$k=1$时A_1是对角占优的$\hat{M}$矩阵。现假定A_k是对角占优的$\hat{M}$矩阵，我们来证A_{k+1}亦是对角占优的$\hat{M}$矩阵。由于A_{k+1}与A_k的前k行完全相同，而A_{k+1}又具有(6.7)所示的形状，故只需证它的右下角的$(n-k)\times(n-k)$子矩阵

$$A_{22}^{(k+1)}=\begin{bmatrix}a_{k+1,k+1}^{(k+1)} & a_{k+1,k+2}^{(k+1)} & \cdots & a_{k+1,n}^{(k+1)}\\ a_{k+2,k+1}^{(k+1)} & a_{k+2,k+2}^{(k+1)} & \cdots & a_{k+2,n}^{(k+1)}\\ \cdots & \cdots & \cdots & \cdots\\ a_{n,k+1}^{(k+1)} & a_{n,k+2}^{(k+1)} & \cdots & a_{nn}^{(k+1)}\end{bmatrix}$$

是对角占优的$\hat{M}$矩阵即可。

比较(6.10)两边的元素，并注意到l_k, a_k和R_k的定义，就有

$$a_{ij}^{(k+1)}=\begin{cases}a_{ij}^{(k)}-l_{ik}a_{kj}^{(k)}, & (i,j)\in\mathscr{T},\\ 0, & (i,j)\notin\mathscr{T}\text{且}i\neq j,\\ a_{ii}^{(k)}-l_{ik}a_{ki}^{(k)}+\omega\sum\limits_{p\in\mathscr{S}(k,i)}(a_{ip}^{(k)}-l_{ik}a_{kp}^{(k)}), & i=j\end{cases}$$

$$(i,j=k+1,k+2,\cdots,n). \tag{6.15}$$

由归纳法假定知，$a_{ij}^{(k)}\leqslant 0$，$i\neq j$，$l_{ik}=a_{ik}^{(k)}/a_{kk}^{(k)}\leqslant 0$，从而由(6.15)立即知

$$a_{ij}^{(k+1)}=a_{ij}^{(k)}-l_{ik}a_{kj}^{(k)}\leqslant a_{ij}^{(k)}\leqslant 0,\qquad (i,j)\in\mathscr{T},$$

$$a_{ij}^{(k+1)}=0=a_{ij}^{(k)},\qquad (i,j)\notin\mathscr{T}\text{ 且 }i\neq j,$$

即(6.14a)成立。

再利用归纳法假定A_k是对角占优的$\hat{M}$矩阵和(6.15)，即知对每个i，$k+1\leqslant i\leqslant n$，有

$$S_i^{(k+1)}=\sum_{i=k+1}^{n}a_{ii}^{(k+1)}$$

$$= \sum_{\substack{j=k+1 \\ (i,j)\in \mathscr{S}}}^{n} (a_{ij}^{(k)} - l_{ik}a_{kj}^{(k)}) + (a_{ii}^{(k)} - l_{ik}a_{ki}^{(k)})$$

$$+ \omega \sum_{p\in \mathscr{S}(k,i)} (a_{ip}^{(k)} - l_{ik}a_{kp}^{(k)})$$

$$\geqslant \sum_{j=k+1}^{n} (a_{ij}^{(k)} - l_{ik}a_{kj}^{(k)})$$

$$= \sum_{j=k+1}^{n} a_{ij}^{(k)} - l_{ik} \sum_{j=k+1}^{n} a_{kj}^{(k)}$$

$$= (S_i^{(k)} - a_{ik}^{(k)}) - l_{ik}(S_k^{(k)} - a_{kk}^{(k)})$$

$$= S_i^{(k)} - l_{ik}S_k^{(k)} - a_{ik}^{(k)} + \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \cdot a_{kk}^{(k)}$$

$$= S_i^{(k)} - l_{ik}S_k^{(k)} \geqslant S_i^{(k)} \geqslant 0,$$

即(6.14b)成立。再结合(6.14a)即有

$$a_{ii}^{(k+1)} - \sum_{\substack{j=k+1 \\ j\neq i}}^{n} |a_{ij}^{(k+1)}| = \sum_{i=k+1}^{n} a_{ij}^{(k+1)} = S_i^{(k+1)} \geqslant 0,$$

$$i = k+1, \cdots, n.$$

由此即知 $A_{22}^{(k+1)}$ 是对角占优矩阵，且有

$$a_{nn}^{(k+1)} \geqslant \sum_{i=k+1}^{n-1} |a_{nj}^{(k+1)}| \geqslant 0,$$

$$a_{ii}^{(k+1)} \geqslant \sum_{\substack{i=k+1 \\ j\neq i}}^{n} |a_{ji}^{(k+1)}| \geqslant |a_{in(i)}^{(k+1)}|$$

$$\geqslant |a_{in(i)}| > 0 \quad (i = k+1, \cdots, n-1),$$

其中 $n(i)$ 是 A 的第 i 行最后一个非零元素所在的列。

此外，由(6.15)和归纳法假定亦有

$$a_{ii}^{(k+1)} = a_{ii}^{(k)} - l_{ik}a_{kj}^{(k)} + \omega \sum_{p\in \mathscr{S}(k,i)} (a_{ip}^{(k)} - l_{ik}a_{kp}^{(k)})$$

$\leqslant a_{ii}^{(k)}(i=k+1,\cdots,n)$。

因此有(6.14c)和(6.14d)成立，而且 $A_{22}^{(k+1)}$ 是 $\hat{M}$ 矩阵。

推论6.1 设 $A\in\mathbb{R}^{n\times n}$ 是对角占优的$\hat{M}$矩阵。则对任意的指标集$\mathscr{T}(\mathscr{T}_n\supset\mathscr{T}\supset\mathscr{D}_A)$和任意的参数$\omega(0\leqslant\omega\leqslant1)$，都存在$A$关于$\mathscr{T}$和$\omega$的松弛不完全LU分解。

这里需特别指出的一点是，虽然在定理6.1的条件下保证了松弛不完全LU分解的存在性，但这并不能保证分解所得到的U是非奇异的(参见例6.1)，即可能有$a_{nn}^{(n)}=0$。

对于给定的矩阵$A=[a_{ij}]\in\mathbb{R}^{n\times n}$，指标集$\mathscr{T}$，$\mathscr{T}_n\supset\mathscr{T}\supset\mathscr{D}_A$，和参数$\omega$，$0\leqslant\omega\leqslant1$，在计算机上求$A$的RILU分解时，可将$L$存放在$A$的下三角部分，$U$存放在$A$的上三角部分，而$R$一般并不需要，因而不保存它。这样，前面介绍的消去修正法就可归纳成如下的实用算法。

算法6.1

(1) 输入$A,\omega,\widetilde{\mathscr{T}}=\mathscr{J}\cup\{(i,i):\ i=1,2,\cdots,n\}$；$r:=1$。

(2) $d:=a_{rr}$，$i:=r+1$。

(3) 如果$(i,r)\in\widetilde{\mathscr{T}}$且$a_{ir}\neq0$，则转步(4)；否则，转步(8)。

(4) $e:=a_{ir}/a$，$a_{ir}:=e$，$j:=r+1$。

(5) 如果$a_{rj}\neq0$，则转步(6)；否则转步(7)。

(6) 如果$(i,j)\in\widetilde{\mathscr{T}}$，则

$$a_{ij}:=a_{ij}-ea_{rj};$$

否则
$$a_{ii}:=a_{ii}-\omega ea_{rj}.$$

(7) 如果$j<n$，则$j:=j+1$，转步(5)；否则转步(8)。

(8) 如果$i<n$，则$i:=i+1$，转步(3)；否则转步(9)。

(9) 如果$r<n-1$，则$r:=r+1$转步(2)；否则输出有关信息，结束。

6.2 松弛不完全Cholesky分解

现在我们再来考虑A为对称正定的情形。此时，我们亦假定指

标集$\mathscr{T}$也是对称的，即若$(i,j)\in\mathscr{T}$，则必有$(j,i)\in\mathscr{T}$。对应于非对称情形的 RILU 分解，我们可以考虑A关于$\mathscr{T}$和ω的**松弛不完全 Cholesky 分解**(简称 RIC 分解)：

$$A = LDL^{\mathrm{T}} + R, \tag{6.16}$$

其中L是单位下三角矩阵且满足(6.3b)，R是剩余矩阵并满足(6.3d)和(6.3e)，D为对角矩阵。易知，此时R亦是对称的，而且亦可采用前面所讲的消去修正法来实现分解(6.16)，即先用消去修正法求A的 RILU 分解

$$A = LU + R,$$

然后令$D=\mathrm{diag}(u_{11},\cdots,u_{nn})$为$U$的对角元素作成的对角矩阵，便得到了(6.16)中的L,D和R。

大家知道，我们求分解式(6.16)的主要目的是，希望能用$M=LDL^{\mathrm{T}}$作为预优矩阵，以提高求解方程组(1.1)的效率。而这就需要LDL^{T}是正定的。下面的定理就给出了保证这样得到的LDL^{T}是正定的一个充分条件。

定理6.2 设$A\in\mathbb{R}^{n\times n}$是对称弱严格对角占优的$M$矩阵，则对任意的对称指标集$\mathscr{T}$，$\mathscr{T}_n\supset\mathscr{T}\supset\mathscr{D}_A$，和任意的松弛参数$\omega$，$0\leqslant\omega\leqslant 1$，$A$都有关于$\mathscr{T}$和$\omega$的松弛不完全 Cholesky 分解

$$A = LDL^{\mathrm{T}} + R,$$

且LDL^{T}是对称正定的，即D的每个对角元素都是正数。

证明完全类似于定理6.1，因此留作练习。

在下面的讨论中，我们假定A满足定理 6.2 的条件。此时，当然我们可以在算法6.1的基础上，利用A的对称性给出求A的RIC分解的算法。但我们亦可从Cholesky 分解出发稍加修改求得。这里略去其详细推理过程，只给出实际求D和L的具体算法。

算法6.2

(1) 输入A，$\mathscr{J}$和ω；$j:=1$，$d_i:=0$，$i=1,2,\cdots,n$。

(2) $d_j := d_j + a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2 d_k$，$i := j+1$。

(3) $\beta := a_{ij} - \sum_{k=1}^{j-1} l_{ik} d_k l_{ik}$。

(4) 如果$(i,j) \in \mathscr{T}$，则

$$l_{ij} := \beta;$$

否则

$$d_j := d_j + \omega\beta, \quad d_i := d_i + \omega\beta.$$

(5) 如果$i < n$，则$i := i+1$，转步(3)；否则

$$l_{ij} := l_{ij}/d_j, \quad i = j+1, j+2, \cdots, n.$$

(6) 如果$j < n$，则 $j := j+1$，转步(2)；否则输出有关信息，结束。

由算法6.2算出L和D之后，我们就可以用$M = LDL^{\mathrm{T}}$作为预优矩阵，应用PCG算法5.1求出方程组(1.1)的解，这就是所谓的**不完全分解预优共轭梯度法**。当$\omega = 0$时，就是所谓的ICCG方法；当$\omega = 1$时，就是所谓的MICCG方法、数值试验和实际应用的结果表明，大多数情况下，MICCG方法优于ICCG方法，而且通常的最优松弛参数ω是与1比较靠近的。

6.3 分块不完全 Cholesky 分解

设A为形如

$$A = \begin{bmatrix} D_1 & A_2^{\mathrm{T}} & & 0 \\ A_2 & D_2 & \ddots & \\ & \ddots & \ddots & A_m^{\mathrm{T}} \\ 0 & & A_m & D_m \end{bmatrix} \tag{6.17}$$

的对称块三对角矩阵，其中D_k是$m_k \times m_k$的三对角矩阵，A_k是$m_k \times m_{k-1}$对角阵(即$A_k = [a_{ij}^{(k)}]$满足$a_{ij}^{(k)} = 0$，$i \neq j$)。

当A是正定矩阵时，它有分块 Cholesky 分解：

$$A=(\Sigma+L)\Sigma^{-1}(\Sigma+L^{\mathrm{T}}),\qquad (6.18\mathrm{a})$$

其中

$$L=\begin{bmatrix} 0 & & & 0 \\ A_2 & 0 & & \\ & \ddots & \ddots & \\ 0 & & A_m & 0 \end{bmatrix},\qquad (6.18\mathrm{b})$$

Σ为块对角矩阵，它的第k块Σ_k由下面的公式递推地产生：

$$\Sigma_1=D_1,$$
$$\Sigma_k=D_k-A_k\Sigma_{k-1}^{-1}A_k^{\mathrm{T}},\quad k=2,\cdots,m.\qquad (6.18\mathrm{c})$$

因为三对角矩阵的逆一般是一个稠密的矩阵，所以由(6.18c)产生的Σ_i不再有稀疏性。因此，若用分块Cholesky分解来求系数矩阵为(6.15)的线性方程组(1.1)，则整个计算过程的存储量会大为增加。这就促使我们寻找合适的预优矩阵M应用PCG方法来求解这样的方程组。

一个很自然的想法是，在(6.18c)中用Σ_i^{-1}的稀疏近似矩阵Λ_i来代替Σ_i^{-1}，即

$$\Delta_1=D_1,$$
$$\Delta_i=D_i-A_i\Lambda_iA_i^{\mathrm{T}},\quad i=2,\cdots,n,\qquad (6.19\mathrm{a})$$

其中Λ_i是Δ_{i-1}^{-1}的某种稀疏近似。这样一来，代替分解式(6.18a)，我们得到了A的分块不完全Cholesky分解(简称BIC)分解：

$$A=(\Sigma+L)\Delta^{-1}(\Delta+L^{\mathrm{T}})-R,\qquad (6.19\mathrm{b})$$

其中$\Delta=\mathrm{diag}(\Delta_1,\cdots,\Delta_m)$，$\Delta_i$由(6.19a)确定，$R=\mathrm{diag}(R_1,\cdots,R_m)$，$R_i$由下面的公式递推地产生：

$$R_1=\Delta_1-D_1=0,$$
$$R_i=\Delta_i-D_i+A_i\Delta_{i-1}^{-1}A_i^{\mathrm{T}},\quad i=2,\cdots,m.\qquad (6.19\mathrm{c})$$

从而我们可选取

$$M=(\Delta+L)\Delta^{-1}(\Delta+L^{\mathrm{T}})\qquad (6.20)$$

作为预优矩阵。

现在的问题是：Λ_i应该怎样选取呢？为此，我们先介绍一个

关于对称三对角矩阵求逆的重要定理。

定理6.3 设T是正定对称三对角矩阵，

$$T=\begin{bmatrix} \alpha_1 & \beta_2 & & 0 \\ \beta_2 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_n \\ 0 & & \beta_n & \alpha_n \end{bmatrix}, \quad \beta_i \neq 0, \quad i=2,\cdots,n.$$

则存在两个n维向量$u=(u_1,\cdots,u_n)^{\mathrm T}$和$v=(v_1,\cdots,v_n)^{\mathrm T}$使得

$$T^{-1}=\begin{bmatrix} u_1v_1 & u_1v_2 & \cdots & u_1v_n \\ & u_2v_2 & \cdots & u_2v_n \\ \text{对称} & & \ddots & \vdots \\ & & & u_nv_n \end{bmatrix}, \tag{6.21}$$

而且u_i, v_i可按如下公式递推地产生：

$$\begin{aligned} &u_1=1, \quad u_2=-\alpha_1/\beta_2 \\ &u_{i+1}=-\left(\frac{\beta_i}{\beta_{i+1}}u_{i-1}+\frac{\alpha_i}{\beta_{i+1}}u_i\right), \quad i=2,\cdots,n-1, \\ &v_n=\frac{1}{\beta_n u_{n-1}+\alpha_n u_n}, \\ &v_{n-1}=-\frac{\alpha_n}{\beta_n}v_n, \\ &v_j=-\left(\frac{\alpha_{j+1}}{\beta_{j+1}}v_{j+1}+\frac{\beta_{j+2}}{\beta_{j+1}}v_{j+2}\right), \quad j=n-2,\cdots,1. \end{aligned} \tag{6.22}$$

证明留作练习。

此外，在一定条件下，可证T^{-1}的三对角部分是T^{-1}的很好的近似(参见本章习题12)。因此，选取Λ_i的一个简单而且有效的方法是选取Λ_i是Δ_{i-1}^{-1}的三对角部分.这样就得到了求形如(6.20)的预优矩阵的有效算法。

算法6.3

(1) 输入$A_2,\cdots,A_m$，$D_1,\cdots,D_m$。

(2) $\Delta_1:=D_1$，$i:=2$。

(3) $\Lambda_i:=\Delta_{i-1}^{-1}$ 的三对角部分(用(6.22)和(6.21)计算)，

$$\Delta_i:=D_i-A_i\Lambda_iA_i^T.$$

(4) 如果 $i<n$，则 $i:=i+1$，转(3)；否则输出有关信息，结束。

关于这一算法的可行性问题，可参看文献[28](Concus 等(1985))，这里不再详述。

此外，在实际应用时，由算法6.3求得 Δ_i 之后，我们还需对 Δ_i 进行 Cholesky 分解

$$\Delta_i=L_iL_i^T.$$

然后用

$$M=\begin{bmatrix} L_1 & & & 0 \\ W_2 & L_2 & & \\ & \ddots & \ddots & \\ 0 & & W_m & L_m \end{bmatrix}\begin{bmatrix} L_1^T & W_2^T & & 0 \\ & L_2^T & \ddots & \\ & & \ddots & W_m^T \\ 0 & & & L_m^T \end{bmatrix}$$

作为预优矩阵来应用PCG 方法，其中

$$W_i=A_iL_{i-1}^{-T},\quad i=2,\cdots,n.$$

当然，在实际计算时，也并不需要具体地计算出 W_i，只需以因子形式保存W_i即可，这是由于实际计算时只涉及求解形如 $Mz=r$ 的方程组之故。

对于二维偏微分方程的离散方程组，分块不完全分解预优共轭梯度法较点不完全分解预优共轭梯度法效率高，但对于三维的情形数值试验的结果表明其效率不如后者(参见文献[67])。

§7 求解非正定线性方程组的共轭梯度法

这一节，我们来考虑如何将前面几节所介绍的关于求解对称正定线性方程组的共轭梯度法推广到一般的线性方程组

$$Ax = b, \tag{7.1}$$

这里仅假定A是一个已知的$n \times n$非奇异实矩阵，b是一已知n维向量，x是待求的n维未知向量。

7.1 正规化方法

应用CG法到方程组(7.1)的一个简单易行的方法就是利用CG法求(7.1)的正规化方程组

$$A^{\mathrm{T}}Ax = A^{\mathrm{T}}b \tag{7.2}$$

的解，其具体算法如下：

算法7.1

(1) 输入A, b和x_0；

$x := x_0$，$r := b - Ax_0$，$p := A^{\mathrm{T}}r$，

$\rho_0 := p^{\mathrm{T}}p$，$k := 0$。

(2) $w := Ap$，$\alpha := \rho_k / w^{\mathrm{T}}w$，

$x := x + \alpha p$，$r := r - \alpha Ap$，

$v := A^{\mathrm{T}}r$，$\rho_{k+1} := v^{\mathrm{T}}v$，

$\beta_k := \rho_{k+1}/\rho_k$，$p := v + \beta_k p$。

(3) 如果$\rho_{k+1} < \rho_0\varepsilon$，则输出$x$，结束；否则$k := k+1$，转步(2)。

这一方法作为求解大型稀疏非正定线性方程组的一种方法，仍具有对称正定情形的一些优点．例如：它可充分利用A的稀疏性；有利于并行化；在没有误差的情况下，可在有限步之内得到(7.1)的精确解等．但由于$A^{\mathrm{T}}A$的条件数是A的条件数的平方，因此当A是病态时，这一方法收敛速度可能变得非常慢．提高收敛速度的一条重要途径就是将对称情形的预优技术推广到非正定情形．一个自然的作法就是应用算法7.1到与(7.1)等价的方程组

$$(L^{-1}AU^{-1})\tilde{x} = \tilde{b}. \tag{7.3}$$

其中$\tilde{x} = Ux$，$\tilde{b} = L^{-1}b$，U和L是需适当选取的非奇异矩阵，目

的在于使$L^{-1}AU^{-1}$的条件数尽可能的小。记$H=LL^{\mathrm{T}}$，$G=U^{\mathrm{T}}U$，可得具体算法如下。

算法7.2

(1) 输入 A,b,H,G 和 x_0；

$x:=x_0$，$r:=b-Ax$，$g:=H^{-1}r$，

$z:=G^{-1}A^{\mathrm{T}}g$，$p:=z$，$\rho_0:=g^{\mathrm{T}}Az$，

$k:=0$。

(2) $g:=H^{-1}Ap$，$\alpha_k:=\rho_k/g^{\mathrm{T}}Ap$，

$x:=x+\alpha_k p$，$r:=r-\alpha_k Ap$，

$g:=H^{-1}r$，$z:=G^{-1}A^{\mathrm{T}}g$，$\rho_{k+1}:=g^{\mathrm{T}}Az$，

$\beta_k:=\rho_{k+1}/\rho_k$，$p:=z+\beta_k p$。

(3) 如果$\rho_{k+1}<\rho_0\varepsilon$，则输出$x$，结束；否则$k:=k+1$，转步(2)。

从上述算法容易看出，每迭代一次必须解形如$Hg=r$和$Gz=g$的两个方程组。这就要求我们在选择L和U时，必须保证这样的方程组容易求解，且L和U应具有一定的稀疏性，否则将毫无意义。上节所介绍的不完全LU分解是解决预优矩阵L和U的选取问题的一种行之有效的方法。

7.2 广义共轭剩余法

正规化方法虽然简单易行，但其收敛速度依赖于A的条件数的平方，当A的条件数很大时，收敛速度变得非常缓慢。因此，近年来人们一直致力于寻找其收敛速度依赖于A的条件数而不是其平方的有效方法，并得到了不少各具千秋的方法。这里，我们介绍其中一种较为有效的方法——广义共轭剩余法。

这类方法的基本思想是，每次迭代在Krylov子空间$\kappa(A,r_0,j)$上求剩余函数

$$\psi(x)=\|b-A(x_0+x)\|_2$$

的极小点，代替了正规化方法每次迭代在Krylov子空间

$\kappa(A^{\mathrm{T}}A,r_0,j)$上求二次泛函

$$\tilde{\varphi}(x)=\frac{1}{2}(x+x_0)^{\mathrm{T}}A^{\mathrm{T}}A(x+x_0)-A^{\mathrm{T}}b^{\mathrm{T}}(x+x_0)$$

的极小点。从第四节关于 CG 法 的收敛性分析可以看出，这样导出的算法的收敛速度应该依赖于A的条件数而不是其平方。

实现这一思想的具体方法如下：首先任取一初始向量x_0，而后计 算 $r_0=b-Ax_0$，并令 $p_0=r_0$；假定我们由此出发已经进行了 k 步，得到了极小化向量 $x_0,x_1,\cdots,x_k$ 和方向向 量 $p_0,\cdots,p_k$；第 $k+1$步是选取x_{k+1}，使

$$\|b-Ax_{k+1}\|_2=\min_{\alpha}\|b-A(x_k+\alpha p_k)\|_2, \tag{7.4}$$

选取新的下降方向p_{k+1}为剩余向量$r_{k+1}=b-Ax_k$在空间 $\mathrm{span}\{p_0,\cdots,p_k\}$的 $A^{\mathrm{T}}A$ 正交补上的 $A^{\mathrm{T}}A$ 正交投影，即

$$p_{k+1}=r_{k+1}+q_{k+1}, \tag{7.5}$$

其中 $q_{k+1}\in\mathrm{span}\{p_0,\cdots,p_k\}$，且

$$p_{k+1}^{\mathrm{T}}A^{\mathrm{T}}Ap_i=0, \qquad i=0,1,2,\cdots,k;$$

而且优化问题(7.4)可完全类似于优化问题(1.5)求得，p_{k+1}可以应用 Gram-Schmidt 正交化求出。

综上所述，就得到了所谓的**广义共轭剩余法**(简称GCR方法)的迭代公式：

$$\begin{aligned}
&\alpha_k=r_k^{\mathrm{T}}Ap_k/p_k^{\mathrm{T}}A^{\mathrm{T}}Ap_k,\\
&x_{k+1}=x_k+\alpha_kp_k,\\
&r_{k+1}=r_k-\alpha_kAp_k,\\
&\beta_j^{(k)}=-r_{k+1}^{\mathrm{T}}A^{\mathrm{T}}Ap_j/p_j^{\mathrm{T}}A^{\mathrm{T}}Ap_j,\quad j=0,1,\cdots,k,\\
&p_{k+1}=r_{k+1}+\sum_{i=0}^{k}\beta_j^{(k)}p_j,
\end{aligned} \tag{7.6}$$

$$k=0,1,2,\cdots,$$

其中 x_0 为给定的初始向量，$r_0=b-Ax_0$，$p_0=r_0$。

容易证明，当A对称正定时，$\beta_j^{(k)}=0$，$j=0,1,2,\cdots,k-1$；

这时，GCR法就是与共轭梯度法等价的共轭剩余法（通常简称为CR法）。当A不是正定矩阵时，可能出现$r_k \neq 0$，但$p_k = 0$的情形；此时，在还没有求得方程组(7.1)的解的情况下，迭代就已中断。保证这种情况不出现的一个充分条件就是A有正定的对称部分，即

$$M = \frac{1}{2}(A + A^{\mathrm{T}}) \tag{7.7}$$

是正定的。因此，在下面的讨论中我们总假定A有正定的对称部分。

定理7.1 设A有正定的对称部分，且假定$\{x_i\}$，$\{r_i\}$和$\{p_i\}$是由GCR迭代产生的。则

(1) $p_i^{\mathrm{T}} A^{\mathrm{T}} A p_j = 0$，$i \neq j$；

(2) $r_i^{\mathrm{T}} A p_j = 0$，$i > j$；

(3) $r_i^{\mathrm{T}} A p_i = r_i^{\mathrm{T}} A r_i$；

(4) $r_i^{\mathrm{T}} A r_j = 0$，$i > j$；

(5) $\mathrm{span}\{p_0, \cdots, p_i\} = \mathrm{span}\{r_0, \cdots, r_i\}$
$= \mathrm{span}\{r_0, Ar_0, \cdots, A^i r_0\}$；

(6) 如果$r_i \neq 0$，则$p_i \neq 0$；

(7) x_{i+1}满足

$\|b - Ax_{i+1}\|_2 = \min\{\|b - A(x_0 + x)\|_2 : x \in \kappa(A, r_0, i+1)\}$

这一定理的证明完全类似于定理3.1的证明，因此这里不再赘述，作为练习请读者自己将其补出。这里需指出的一点是，这一定理的结论除(6)用到A有正定的对称部分外，其余各条都不需这一假定。

从这一定理可以看出GCR迭代亦有类似于CG法的一些优点，而且作为迭代法由(7)易得

定理7.2 假定r_k是由GCR迭代第k步产生的剩余向量，则有

$$\|r_k\|_2 \leqslant \min_{q_k \in \mathscr{P}_k^{(0)}} \|q_k(A)\|_2 \|r_0\|_2, \tag{7.8}$$

其中$\mathscr{P}_k^{(0)}$表示常数项为 1 且次数不超过 k 的实系数多项式全体。

利用(7.8)，通过选取各种不同的多项式q_k，可以得到各种各样的收敛上界的估计。例如，取 $q_k(t)=(1-\alpha t)^k$，$\alpha \in \mathbb{R}$，可证

$$\begin{aligned}\|r_k\|_2 &\leqslant \min\{\|(I-\alpha A)^k\|_2 \|r_0\|_2 :\ \alpha \in \mathbb{R}\} \\ &\leqslant \left[1-\frac{\lambda_{\min}^2(M)}{\lambda_{\max}(A^{\mathrm{T}}A)}\right]^{k/2} \|r_0\|_2, \end{aligned} \tag{7.9}$$

其中M如(7.7)所定义，$\lambda_{\min}(M)$表示M的最小特征值，$\lambda_{\max}(A^{\mathrm{T}}A)$表示$A^{\mathrm{T}}A$的最大特征值。

此外，当A是正规矩阵时，(7.8)即为

$$\|r_k\|_2 \leqslant \min_{q_k \in \mathscr{P}_k^{(0)}} \max_{\lambda \in \lambda(A)} |q_k(\lambda)| \|r_0\|_2. \tag{7.10}$$

从 GCR 的迭代公式(7.6)可以看出，p_{k+1}的计算工作量是相当大的，而且还需保留前面所得到的所有方向向量 $p_i(i=0,1,\cdots,k)$，随着迭代次数的增加，这将要占用大量的内存空间。因此，在实际使用时，为减少工作量节约内存，通常采用如下两种措施：一是选择一个适当的正数 s，只要求p_{k+1}与$p_k,\cdots,p_{k-s+1}$保持$A^{\mathrm{T}}A$正交，即在(7.6)中只计算$\beta_k^{(k)},\cdots,\beta_{k-s+1}^{(k)}$，而令 $\beta_1^{(k)},\cdots,\beta_{k-s}^{(k)}$都为零，这就是所谓的 Orthomin($s$)算法；二是对适当选择的正数 s 和给定的初始向量 x_0，利用(7.6)迭代 s 步之后，再以 x_s作为初始向量重新用(7.6)迭代 s 步，这样周而复始的进行，就是所谓的GCR(s)算法。下面我们只给出 GCR(s)算法的具体内容，而 Orthomin(s)算法请读者作为练习自己总结出来。

算法7.3

(1) 输入 A,b,x_0 和 s；$x:=x_0$。

(2) $r:=b-Ax$，$p_0:=r$，$\rho:=\|r\|_2$，$k:=0$。

(3) $\alpha := r^{\mathrm{T}} A p_k / p_k^{\mathrm{T}} A^{\mathrm{T}} A p_k$,

$x := x + \alpha p_k$,

$r := r - \alpha A p_k$,

$\beta_j := r^{\mathrm{T}} A^{\mathrm{T}} A p_j / p_j^{\mathrm{T}} A^{\mathrm{T}} A p_j, \quad j = 0, 1, \cdots, k$

$p_{k+1} := r + \sum_{j=0}^{k} \beta_j p_j$.

(4) 如果 $k = s$，则转步(5)；否则 $k := k + 1$，转步(3)。

(5) 如果 $\|r\|_2 < \rho\varepsilon$，则输出 x 停机；否则转步(2)。

究竟 s 取多大为最好，还没有理论上的结果，但GCR(s)和Orthomin(s)算法为很多文献所采用,大都选取$s \leqslant 20$,效果都很好。

习　题

1．设 k 维超平面 π_k 的法向为 $p_1, \cdots, p_{n-k}$。试证：p 垂直于 π_k 的充要条件是 $p \in \mathrm{span}\{p_1, \cdots, p_{n-k}\}$。

2．设$\tilde{\pi}_{n-k}$是 k 维超平面π_k的共轭超平面，且π_k是由$u_1, \cdots, u_k$张成的。证明：$Au_1, Au_2, \cdots, Au_k$就是$\tilde{\pi}_{n-k}$的法向，且有

$$\tilde{\pi}_{n-k}: \; u_i^{\mathrm{T}}(Ax - b) = 0, \quad i = 1, 2, \cdots, k.$$

3．考虑系数矩阵为

$$A = \begin{bmatrix} 4 & -2 & -1 & 0 \\ -2 & 4 & 0 & -2 \\ -1 & 0 & 4 & -1 \\ 0 & -2 & -1 & 4 \end{bmatrix}$$

的线性方程组。证明：

(1) 共轭梯度法应用到此方程组上，对任意的初值至多迭代3次就可得到方程组的精确解；

(2) 如果初始向量x_0使得剩余向量为 $r_0 = (1, 1, -2, -1)^{\mathrm{T}}$，则此时只需迭代一次就可得到方程组的精确解。

4．设 A 为对角元素均为 2，次对角元素均为 -1 的 n 阶对称三对角矩阵，L 为如下的二对角矩阵

$$L=\begin{bmatrix} 1 & & & & 0 \\ -1 & 1 & & & \\ & -1 & \ddots & & \\ 0 & & \ddots & 1 & \\ & & & -1 & 1 \end{bmatrix}.$$

(1) 证明：$\tilde{A}=L^{-1}AL^{-T}$ 的特征值为 $1(n-1$重$)$和 $n+1($单重$)$；

(2) 证明：对系数矩阵为上述 A 的线性方程组，可选取预优矩阵 M，使得PCG法最多只需迭代两次就收敛。

5. 试证：如果系数矩阵 A 至多有 l 个互不相同的特征值，则共轭梯度法至多 l 步就可得到方程组 $Ax=b$的精确解。

6. 证明：如果 $A=[a_{ij}]\in\mathbb{R}^{n\times n}$ 是严格对角占优的，则 A 有 RILU 分解。

7. 举例说明正定矩阵不一定有 RIC 分解。

8. 证明：对任意的$A\in\mathbb{R}^{n\times n}$，有 $\lambda(M)\subset\{\lambda:\ |\lambda|\leqslant\|A\|_2\}$，其中$M=(A^T+A)/2$为 A 的对称部分。

9. 证明：当 A 正定对称时，广义共轭剩余法迭代公式中的 $\beta_j^{(k)}=0(j=0,1,\cdots,k-1)$。并说明 A 不对称时，为什么这一事实不真。

10. 举例说明，当 A 的对称部分不正定时，可在广义共轭剩余法中出现 $r_i\neq0$，而$p_i=0$的情形。

11. 计算矩阵

$$A=\begin{bmatrix} 4 & -1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 4 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 4 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 4 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 4 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 3 & -1 \\ 0 & 0 & 0 & 0 & 1 & -1 & 3 \end{bmatrix}$$

的不完全分块Cholesky分解。

12．设严格对角占优的实对称三对角阵T的元素α_i和β_i满足：

$$\alpha_i>0,\ 1\leqslant i\leqslant n;\quad \beta_i<0,\ 2\leqslant i\leqslant n.$$

试证：

（1）由(6.22)产生的$\{u_i\}_1^n$和$\{v_i\}_1^n$满足

$$1=u_1<u_2<\cdots<u_n,$$

$$v_1>v_2>\cdots>v_n>0;$$

（2）$0<(T^{-1})_{ij}\leqslant(T^{-1})_{ii}\rho^{i-j}$，$1\leqslant j<i\leqslant n$，其中$(T^{-1})_{ij}$表示$T^{-1}$的第$(i,j)$位置上的元素，$\rho=\max\limits_{2<j<n}\{-\beta_{j+1}/(\alpha_j+\beta_j)\}<1$。

13．试给出一个不完全分块Cholesky分解存在的充分条件。

第六章　最小二乘问题的数值解法

§1　最小二乘解的数学性质

1.1　最小二乘解的特征

设 $A\in\mathbb{R}^{m\times n}(m>n)$，$b\in\mathbb{R}^m$. 所谓线性最小二乘问题(简称LS问题)，是指求 $x\in\mathbb{R}^n$ 使得

$$\|Ax-b\|_2=\min\{\|Av-b\|_2:\ v\in\mathbb{R}^n\}. \tag{1.1}$$

记

$$\mathscr{X}_{\mathrm{LS}}=\{x\in\mathbb{R}^n:\ x\text{ 满足(1.1)}\}. \tag{1.2}$$

则称 $\mathscr{X}_{\mathrm{LS}}$ 是LS问题(1.1)的**解集**；$\mathscr{X}_{\mathrm{LS}}$ 中2范数最小者称作**最小范数解**，记作 x_{LS}，即

$$\|x_{\mathrm{LS}}\|_2=\min\{\|x\|_2:\ x\in\mathscr{X}_{\mathrm{LS}}\}.$$

定理1.1　$x\in\mathscr{X}_{\mathrm{LS}}$ 当且仅当 $A^{\mathrm{T}}(Ax-b)=0$.

证明　由于对任意的 $x,y\in\mathbb{R}^n$，有

$$\|b-A(x+y)\|_2^2=\|b-Ax\|_2^2-2y^{\mathrm{T}}A^{\mathrm{T}}(b-Ax)+\|Ay\|_2^2,$$

因此，$x\in\mathscr{X}_{\mathrm{LS}}$ 当且仅当对任意的 $y\in\mathbb{R}^n$ 有

$$\|Ay\|_2^2-2y^{\mathrm{T}}A^{\mathrm{T}}(b-Ax)\geqslant 0. \tag{1.3}$$

易证(1.3)对任意的 $y\in\mathbb{R}^n$ 成立的充分必要条件是

$$A^{\mathrm{T}}(b-Ax)=0. \tag{1.4}$$

事实上，(1.4)成立显然有(1.3)成立；反之，若(1.4)不成立，令

$$y=\varepsilon A^{\mathrm{T}}(b-Ax),$$

其中 $\varepsilon\in\mathbb{R}$，则有

$$\|Ay\|_2-2y^{\mathrm{T}}A^{\mathrm{T}}(b-Ax)$$

$$= \varepsilon^2 \| AA^{\mathrm{T}}(b - Ax) \|_2^2 - 2\varepsilon \| A^{\mathrm{T}}(b - Ax) \|_2^2 < 0$$

对充分小的正数 ε 成立，即(1.3)不成立。证毕。

推论1.1 (1) $\mathscr{X}_{\mathrm{LS}}$ 是凸集；

(2) x_{LS} 是唯一的；

(3) $\mathscr{X}_{\mathrm{LS}} = \{x_{\mathrm{LS}}\}$ 的充分必要条件是 $\mathrm{rank}(A) = n$。

1.2 最小二乘解的一般表示

为了给出最小二乘解的一般表示，我们需要广义逆的一些基本性质。

定义1.1 设 $A \in \mathbb{R}^{m \times n}$。如果 $X \in \mathbb{R}^{n \times m}$ 满足：

(1) $AXA = A$,

(2) $XAX = X$,

(3) $(AX)^{\mathrm{T}} = AX$,

(4) $(XA)^{\mathrm{T}} = XA$,

则称 X 是 A 的 **Moore-Penrose 广义逆**，记作 $A^{\dagger}$。

可以证明：$A^{\dagger}$ 是由 A 唯一确定的，而且若 A 的SVD 分解为

$$A = U \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} V^{\mathrm{T}},$$

其中 U 和 V 分别为 m 阶和 n 阶正交矩阵，$\Sigma_r = \mathrm{diag}(\sigma_1, \cdots, \sigma_r)$，$\sigma_1 \geqslant \cdots \geqslant \sigma_r > 0$，则

$$A^{\dagger} = V \begin{bmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^{\mathrm{T}}. \tag{1.5}$$

广义逆的另一个重要性质是，它可给出由 A 确定的四个基本空间上的正交投影算子的表达式：

$$P_{\mathscr{R}(A)} = AA^{\dagger}, \quad P_{\mathscr{N}(A^{\mathrm{T}})} = I - AA^{\dagger}, \tag{1.6a}$$

$$P_{\mathscr{R}(A^{\mathrm{T}})} = A^{\dagger}A, \quad P_{\mathscr{N}(A)} = I - A^{\dagger}A. \tag{1.6b}$$

由于篇幅所限，有关广义逆的详细讨论这里不再给出，有兴趣的读者可参阅文献[4]。

利用广义逆的上述基本性质，我们可给出最小二乘问题之解的一般表示。

定理1.2 LS 问题(1.1)的解由

$$x = A^{\dagger}b + (I - A^{\dagger}A)z \tag{1.7}$$

给出，其中 z 表示 $\mathbb{R}^n$ 中的任一向量；而且其唯一的最小范数解由

$$x_{\mathrm{LS}} = A^{\dagger}b \tag{1.8}$$

给出。

证明 由定理 1.1 知，LS 问题(1.1)的解可由正规化方程组

$$A^{\mathrm{T}}Ax = A^{\mathrm{T}}b \tag{1,9}$$

给出。

利用广义逆的定义，直接验证可知 $A^{\dagger}b$ 是方程组(1.9)的一个解。

此外，利用奇异值分解定理易证 $\mathcal{N}(A^{\mathrm{T}}A) = \mathcal{N}(A)$，而

$$\mathcal{N}(A) = \{P_{\mathcal{N}(A)}z:\ z \in \mathbb{R}^n\} = \{(I - A^{\dagger}A)z:\ z \in \mathbb{R}^n\},$$

因此，由线性方程组的基本理论知，方程组(1.9)的解由

$$x = A^{\dagger}b + (I - A^{\dagger}A)z, \quad z \in \mathbb{R}^n$$

给出，即 LS 问题(1.1)之解可由(1.7)给出。

另外，注意到

$$[(I - A^{\dagger}A)z]^{\mathrm{T}}A^{\dagger}b = z^{\mathrm{T}}(I - A^{\dagger}A)A^{\dagger}b = 0,$$

就有

$$\|x\|_2^2 = \|A^{\dagger}b\|_2^2 + \|(I - A^{\dagger}A)z\|_2^2 \geq \|A^{\dagger}b\|_2^2,$$

对一切的 $z \in \mathbb{R}^n$ 成立，从而有 $x_{\mathrm{LS}} = A^{\dagger}b$。证毕。

1.3 最小二乘解的扰动分析

假定 $A, \delta A \in \mathbb{R}^{m \times n}$ 和 $b, \delta b \in \mathbb{R}^m$；并假设 x 和 $x + \delta x \in \mathbb{R}^m$分别是 LS 问题

$$\|Ax-b\|_2=\min\{\|Av-b\|_2:\ v\in\mathbf{R}^n\}$$

和

$$\begin{aligned}&\|(A+\delta A)(x+\delta x)-(b+\delta b)\|_2\\&\quad=\min\{\|(A+\delta A)v-(b+\delta b)\|_2:\ v\in\mathbf{R}^n\}\end{aligned}$$

的最小范数解，即

$$x=A^{\dagger}b,\quad x+\delta x=(A+\delta A)^{\dagger}(b+\delta b).$$

现在，我们来考虑 δA 和 δb 的大小对 δx 的影响。

首先，需要指出的一点是：广义逆与通常的矩阵的逆不同，它是不连续的，即当 δA 趋向于零时，$(A+\delta A)^{\dagger}$ 不一定趋向于 $A^{\dagger}$。这样的例子很容易举出，例如，设

$$A=\begin{bmatrix}1&0\\0&0\\0&0\end{bmatrix},\quad \delta A=\begin{bmatrix}0&0\\0&\varepsilon\\0&0\end{bmatrix},\quad \varepsilon>0.$$

则有

$$A^{\dagger}=\begin{bmatrix}1&0&0\\0&0&0\end{bmatrix},\quad (A+\delta A)^{\dagger}=\begin{bmatrix}1&0&0\\0&\dfrac{1}{\varepsilon}&0\end{bmatrix},$$

而且

$$\|A^{\dagger}-(A+\delta A)^{\dagger}\|_2=\frac{1}{\varepsilon}\to\infty\quad(\varepsilon\to 0).$$

广义逆的不连续性，使得关于LS问题之解的扰动分析变得复杂化。值得欣慰的是，Stewart (1969)通过研究广义逆矩阵的扰动界，揭示了广义逆的连续性与保秩扰动之间的内在联系。他证明了：

$$\lim_{\|\delta A\|_2\to 0}(A+\delta A)^{\dagger}=A^{\dagger}$$

的充分必要条件是，这些 δA 在充分靠近零时，有 $\mathrm{rank}(A+\delta A)=\mathrm{rank}(A)$。因此，我们自然应在 δA 不改变 A 的秩的前提下，来考虑 δx 的大小与 δA 和 δb 之间的关系。

为了下面叙述简单起见，我们引进记号：

$$\tilde{A}=A+\delta A,\quad \tilde{x}=x+\delta x,\quad \tilde{b}=b+\delta b,$$

$\varepsilon_A = \|\delta A\|_2/\|A\|_2$, $\kappa = \kappa_1(A) = \|A\|_2\|A^\dagger\|_2$, $\eta = \kappa\varepsilon_A$.

定理1.3 如果 $\text{rank}(\tilde{A}) = \text{rank}(A)$，且 $\eta<1$，则

$$\|\delta x\|_2 \leqslant \frac{\kappa}{1-\eta}\left(\varepsilon_A\|x\|_2 + \frac{\|\delta b\|_2}{\|A\|_2} + \varepsilon_A\kappa\frac{\|r\|_2}{\|A\|_2}\right) + \varepsilon_A\kappa\|x\|_2, \tag{1.10}$$

其中 $r = b - Ax$。

证明 由于

$$\begin{aligned}\delta x = \tilde{x} - x &= \tilde{A}^\dagger\tilde{b} - x \\ &= \tilde{A}^\dagger(Ax + r + \delta b) - x \\ &= \tilde{A}^\dagger(-\delta Ax + \delta b) + \tilde{A}^\dagger r + (\tilde{A}^\dagger\tilde{A} - I)x \\ &= \tilde{A}^\dagger(\delta b - \delta Ax) + \tilde{A}^\dagger r - P_{\mathscr{N}(\tilde{A})}x,\end{aligned} \tag{1.11}$$

因此，要给出 $\|\delta x\|_2$ 的上界估计，只需依次给出(1.11)右边三项之范数的上界估计即可。

$$\begin{aligned}\|\tilde{A}^\dagger(\delta b - \delta Ax)\|_2 &\leqslant \|\tilde{A}^\dagger\|_2(\|\delta b\|_2 + \|\delta A\|_2\|x\|_2) \\ &\leqslant \frac{\|A^\dagger\|_2}{1-\eta}(\|\delta b\|_2 + \|\delta A\|_2\|x\|_2) \\ &= \frac{\kappa}{1-\eta}\left(\frac{\|\delta b\|_2}{\|A\|_2} + \varepsilon_A\|x\|_2\right).\end{aligned} \tag{1.12}$$

上述估计中，第二个不等式用到了

$$\|\tilde{A}^\dagger\|_2 \leqslant \frac{1}{1-\eta}\|A^\dagger\|_2. \tag{1.13}$$

这一不等式的证明留给读者。

注意到 $r \in \mathscr{N}(A^{\mathrm{T}})$ 和 $\tilde{A}^\dagger P_{\mathscr{R}(\tilde{A})} = \tilde{A}^\dagger$，就有

$$\tilde{A}^\dagger r = \tilde{A}^\dagger P_{\mathscr{R}(\tilde{A})} P_{\mathscr{N}(A^{\mathrm{T}})} r; \tag{1.14}$$

而利用第一章习题 5 的结论，有

$$
\begin{aligned}
\|P_{\mathcal{R}(\tilde{A})} P_{\mathcal{N}(A^{\mathrm{T}})}\|_2 &= \|P_{\mathcal{R}(\tilde{A})} - P_{\mathcal{R}(A)}\|_2 = \|P_{\mathcal{R}(A)} P_{\mathcal{N}(\tilde{A}^{\mathrm{T}})}\|_2 \\
&= \|P_{\mathcal{N}(\tilde{A}^{\mathrm{T}})} P_{\mathcal{R}(A)}\|_2 = \|(I - \tilde{A}\tilde{A}^{\dagger})\tilde{A}\tilde{A}^{\dagger}\|_2 \\
&= \|(I - \tilde{A}\tilde{A}^{\dagger})(A - \tilde{A})A^{\dagger}\|_2 \\
&= \|P_{\mathcal{N}(\tilde{A}^{\mathrm{T}})} \delta A\, A^{\dagger}\|_2 \\
&\leqslant \|\delta A\|_2 \|A^{\dagger}\|_2 \\
&= \kappa \varepsilon_A ;
\end{aligned}
\tag{1.15}
$$

因此，在(1.14)两边取 2 范数，并利用(1.15)和(1.13)，可得

$$
\begin{aligned}
\|\tilde{A}^{\dagger} r\|_2 &\leqslant \|\tilde{A}^{\dagger}\|_2 \|P_{\mathcal{R}(\tilde{A})} P_{\mathcal{N}(A^{\mathrm{T}})}\|_2 \|r\|_2 \\
&\leqslant \frac{\|A^{\dagger}\|_2}{1-\eta} \kappa \varepsilon_A \|r\|_2 \\
&= \frac{\kappa^2}{1-\eta} \varepsilon_A \cdot \frac{\|r\|_2}{\|A\|_2} .
\end{aligned}
\tag{1.16}
$$

由于

$$
P_{\mathcal{R}(A^{\mathrm{T}})} x = (A^{\dagger} A)(A^{\dagger} b) = A^{\dagger} b = x ,
$$

所以

$$
\begin{aligned}
\|P_{\mathcal{N}(\tilde{A})} x\|_2 &= \|P_{\mathcal{N}(\tilde{A})} P_{\mathcal{R}(A^{\mathrm{T}})} x\|_2 \\
&\leqslant \|P_{\mathcal{N}(\tilde{A})} P_{\mathcal{R}(A^{\mathrm{T}})}\|_2 \|x\|_2 \\
&= \|P_{\mathcal{R}(A^{\mathrm{T}})} P_{\mathcal{N}(\tilde{A})}\|_2 \|x\|_2 \\
&= \|A^{\dagger}(A - \tilde{A}) P_{\mathcal{N}(\tilde{A})}\|_2 \|x\|_2 \\
&\leqslant \|A^{\dagger}\|_2 \|\delta A\|_2 \|x\|_2 \\
&= \kappa \varepsilon_A \|x\|_2 .
\end{aligned}
\tag{1.17}
$$

在(1.11)两边取 2 范数，并将(1.12),(1.16)和(1.17) 代入立即得到(1.10)。证毕。

定理1.3表明，

$$
\kappa = \kappa_2(A) = \|A^{\dagger}\|_2 \|A\|_2
$$

的大小，在一定程度上反映了最小二乘问题之解对扰动的敏感程度。因此，类比于线性方程组的情形，我们称 $\kappa_2(A)$ 为最小二乘问题(1.1)的**条件数**；$\kappa_2(A)$ 很大，就说问题(1.1)是**病态的**；否则就说其是**良态的**。

从定理 1.3 的证明可知，当 $\text{rank}(A)=\text{rank}(\tilde{A})=n$ 时，(1.10)的最后一项可去掉，因这时 $P_{\mathscr{N}(\tilde{A})}=0$。

(1.10)的第三项含有因子 κ^2(条件数的平方项)和 $\|r\|_2/\|A\|_2$，这一项反映了最小二乘问题之剩余向量的大小对解的敏感性的影响；当剩余向量较大时，问题之解的敏感程度主要取决于条件数的平方的大小；当剩余向量很小（以致于第三项可以忽略）时，问题之解的敏感程度就只线性依赖于条件数；特别当 $b\in\mathscr{R}(A)$，即 $r=0$ 时，第三项不出现；当 A 是可逆方阵时，(1.10)就变成线性方程之解的扰动上界估计，与第三章的定理1.2相吻合。

最后需指出的是，虽然(1.10)所给出的估计并非最好，然而其第三项依赖于 κ^2 这一事实是不可改进的，在极特殊的情况下是可以达到的。例如，设

$$A=\begin{bmatrix}1&0\\0&10^{-6}\\0&0\end{bmatrix},\quad \delta A=\begin{bmatrix}0&0\\0&0\\0&10^{-8}\end{bmatrix},\quad b=\begin{bmatrix}1\\0\\1\end{bmatrix},\ \delta b=\begin{bmatrix}0\\0\\0\end{bmatrix}.$$

容易算出

$$x=(1,0)^{\mathrm{T}},\quad \tilde{x}\approx(1,0.9999\times10^4)^{\mathrm{T}},$$

$$\kappa=\kappa_2(A)=10^6,\quad \varepsilon_A=\|\delta A\|_2/\|A\|_2=10^{-8}.$$

此时，

$$\frac{\|\delta x\|_2}{\|x\|_2}=\frac{\|\tilde{x}-x\|_2}{\|x\|_2}\approx0.9999\times10^4\approx10^{12}\times10^{-8}=\kappa^2\varepsilon_A.$$

§2 求解满秩LS问题的数值方法

这一节，我们假定 LS 问题(1.1)中的矩阵 A 是满秩的，即

rank$(A)=n$。此时，LS 问题(1.1)有且仅有唯一的解 x_{LS}，并且连续地依赖于给定的数据 A 和 b。有关求解这类问题的数值方法在一般的大学教课书里都有详尽的讨论，因而这里我们只对其中最基本的方法作一概述性介绍。

2.1 正规化方法

求解 LS 问题之解的最古老而且现在仍常用的方法之一就是正规化方法。这一方法的理论依据就是定理1.1，即将求解 LS 问题(1.1)转化为求解正规化方程组

$$A^{T}Ax=A^{T}b. \tag{2.1}$$

当 rank$(A)=n$ 时，$A^{T}A$ 是对称正定矩阵，因而(2.1) 的唯一解 $x=x_{LS}$ 可用 Cholesky 分解法求得。因此，正规化方法的基本步骤为:

（1）计算 $C=A^{T}A$ 和 $d=A^{T}b$；

（2）计算 C 的 Cholesky 分解 $C=GG^{T}$；

（3）求解三角方程组 $Gy=d$ 和 $G^{T}x=y$。

这一方法的运算量是 $\frac{1}{2}n^{2}\left(m+\frac{1}{3}n\right)$。设 $\hat{x}$ 是利用这一方法求解(2.1)所得到的计算解，x 是它的精确解。则可证（见文献[20]）

$$\|\hat{x}-x\|\leqslant 2.5n^{\frac{3}{2}}\varepsilon\kappa_{2}^{2}(A)\|x\|_{2}. \tag{2.2}$$

换句话说，就是正规化方法的计算结果的精确程度依赖于 A 的条件数的平方。因此，它与下面将要介绍的正交化方法相比，数值稳定性较差。但由于这一方法简单易用，现在仍然经常使用；特别当 $m\gg n$(即方程的个数远远大于未知数的个数)时，这一方法具有一定的优越性，因为这时 $A^{T}A$ 将占用比 A 少得多的存储空间。

2.2 正交化方法

由于矩阵的 2 范数是正交不变的，因此对任意的正交矩阵

$Q\in\mathbb{R}^{m\times m}$，问题(1.1)等价于

$$\|Q^{\mathrm{T}}(Ax-b)\|_2=\min\{\|Q^{\mathrm{T}}(Av-b)\|_2:\ v\in\mathbb{R}^n\}. \tag{2.3}$$

这样，我们就可望通过适当选取正交矩阵 Q，使原问题转化为较容易求解的 LS 问题(2.3)，这就是正交化方法的基本思想。

设 A 有 QR 分解：

$$A=Q\begin{bmatrix}R\\0\end{bmatrix}=Q_1R, \tag{2.4}$$

其中 $Q\in\mathbb{R}^{m\times m}$ 是正交矩阵，Q_1 是 Q 的前 n 列组成的矩阵，即 $Q=[\underset{n}{Q_1},\ \underset{m-n}{Q_2}]$，$R\in\mathbb{R}^{n\times n}$ 是对角线均为正数的上三角矩阵。

现在取(2.3)中的正交矩阵就是分解式(2.4)中的 Q，并记

$$d=Q^{\mathrm{T}}b=\begin{bmatrix}Q_1^{\mathrm{T}}\\Q_2^{\mathrm{T}}\end{bmatrix}b=\begin{bmatrix}d_1\\d_2\end{bmatrix}\begin{matrix}n\\m-n\end{matrix}, \tag{2.5}$$

则有

$$\begin{aligned}\|Q^{\mathrm{T}}(Ax-b)\|_2^2&=\left\|\begin{bmatrix}R\\0\end{bmatrix}x-\begin{bmatrix}d_1\\d_2\end{bmatrix}\right\|_2^2\\&=\|Rx-d_1\|_2^2+\|d_2\|_2^2.\end{aligned} \tag{2.6}$$

由此易知，x 是 LS 问题(1.1)的解当且仅当 x 是方程组 $Rx=d_1$ 的解。这样一来，LS 问题(1.1)的解就可很容易从上三角方程组 $Rx=d_1$ 求得。

综合上面的讨论，可得正交化方法的基本步骤为：

(1) 求分解式(2.4)中的 Q_1 和 R；

(2) 计算 $d_1=Q_1^{\mathrm{T}}b$；

(3) 解方程组 $Rx=d_1$。

由此可见，实现正交化方法的关键在于如何计算分解式(2.4)。这方面的最基本方法有下面三种：

(i) Householder 方法；

(ii) Givens方法;

(iii) 改进的 Gram-Schmit 正交化方法。

这里，我们只对 Householder 方法作一简述，其他两种方法可查阅有关的教课书。

用 Householder 方法计算分解式(2.4)，就是利用 Householder 变换逐步将 A 约化为上三角矩阵。

设我们已确定了 $k-1$ 个 Householder 变换 $H_1,\cdots,H_{k-1}$，使得

$$A_{k-1}=H_{k-1}\cdots H_1A=\begin{bmatrix} A_{11}^{(k-1)} & A_{12}^{(k-1)} \\ 0 & A_{22}^{(k-1)} \end{bmatrix},$$

其中 $A_{11}^{(k-1)}$ 是$(k-1)\times(k-1)$的上三角阵。现记 v_{k-1} 为 $A_{22}^{(k-1)}$ 的第一列。我们的第 k 步是：先确定Householder 变换 $\tilde{H}_k\in\mathbb{R}^{(m-k+1)\times(m-k+1)}$，使得

$$\tilde{H}_k v_{k-1}=r_{kk}e_1,$$

其中 $r_{kk}\in\mathbb{R}$，$e_1=(1,0,\cdots,0)\in\mathbb{R}^{m-k+1}$；然后，我们再计算 $\tilde{H}_kA_{22}^{(k-1)}$。令

$$H_k=\mathrm{diag}(I_{k-1},\tilde{H}_k),$$

则有

$$A_k=H_kA_{k-1}=\begin{bmatrix} A_{11}^{(k-1)} & A_{12}^{(k-1)} \\ 0 & \tilde{H}_kA_{22}^{(k-1)} \end{bmatrix}=\begin{bmatrix} A_{11}^{(k)} & A_{12}^{(k)} \\ 0 & A_{22}^{(k)} \end{bmatrix},$$

其中 $A_{11}^{(k)}$ 是 $k\times k$ 的上三角阵。这样，从 $k=1$ 出发，依次进行 n 次，我们就可将 A 约化为上三角阵。现记

$$Q=(H_n\cdots H_1)^{\mathrm{T}}=H_1H_2\cdots H_n,$$
$$R=A_{11}^{(n)},$$

则有

$$A=Q\begin{bmatrix} R \\ 0 \end{bmatrix}$$

即为所求的分解式(2.4)。

如果我们求分解式(2.4)只是为了求LS问题(1.1)的解，则上述方法所得到的Q不必明确地计算出来。而且H_k也不需保存，只需在每次确定H_k之后将其作用在向量b上即可。这样，便可得到求解LS问题(1.1)的如下算法。

算法2.1

(1) 输入$A=[a_{ij}]\in\mathbb{R}_n^{m\times n}$ $(m>n)$，$b\in\mathbb{R}^m$；$k:=1$。

(2) 确定Householder变换$\tilde{H}_k\in\mathbb{R}^{(m-k+1)\times(m-k+1)}$，使得

$$\tilde{H}_k\begin{bmatrix}a_{kk}\\a_{k+1,k}\\\vdots\\a_{mk}\end{bmatrix}=\begin{bmatrix}r_{kk}\\0\\\vdots\\0\end{bmatrix};\quad a_{kk}:=r_{kk},$$

$$\begin{bmatrix}a_{k,k+1} & \cdots & a_{kn}\\a_{k+1,k+1} & \cdots & a_{k+1,n}\\\cdots\cdots & \cdots\cdots & \cdots\cdots\\a_{m,k+1} & \cdots & a_{mn}\end{bmatrix}:=\tilde{H}_k\begin{bmatrix}a_{k,k+1} & \cdots & a_{kn}\\a_{k+1,k+1} & \cdots & a_{k+1,n}\\\cdots\cdots & \cdots\cdots & \cdots\cdots\\a_{m,k+1} & \cdots & a_{mn}\end{bmatrix},$$

$$(b_k,\cdots,b_m)^{\mathrm{T}}:=\tilde{H}_k(b_k,\cdots,b_m)^{\mathrm{T}}.$$

(3) 如果$k<n$，则$k:=k+1$，转步(2)；否则进行下一步。

(4) 求解方程组

$$\begin{bmatrix}a_{11} & a_{12} & \cdots & a_{1n}\\ & a_{22} & \cdots & a_{2n}\\ 0 & & \ddots & \vdots\\ & & & a_{nn}\end{bmatrix}\begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix}=\begin{bmatrix}b_1\\b_2\\\vdots\\b_n\end{bmatrix}.$$

(5) 输出$x=(x_1,\cdots,x_n)^{\mathrm{T}}$，结束。

这一算法的运算量是$n^2(m-n/3)$，大约是正规化方法的2倍。

设$\tilde{x}$是由算法2.1计算得到的，则$\tilde{x}$满足

$$\begin{aligned}&\|(A+\delta A)\tilde{x}-(b+\delta b)\|_2\\&\quad=\min\{\|(A+\delta A)v-(b+\delta b)\|_2:\ v\in\mathbb{R}^n\},\end{aligned}$$

其中

$$\|\delta A\|_2 \leqslant c_2 \varepsilon n^{1/2} \|A\|_2,$$
$$\|\delta b\|_2 \leqslant c_2 \varepsilon \|b\|_2,$$
$$c_2 = (6m - 3n + 41)n,$$

详见文献[20]。

由此可见，这一算法有良好的数值稳定性；特别当所计算的LS问题是小剩余问题时，计算结果要比正规化方法精确的多，当然，付出的代价也是不容忽视的。

此外，当 $\kappa_2(A) \approx \varepsilon^{-1}$ 时，这一算法可能会出现中断，而当 $\kappa_2(A) \approx \varepsilon^{-1/2}$ 时，正规化方法中的 Cholesky 分解就可能出问题，因此，这一方法比正规化方法的应用范围广。

§3 求解亏秩LS问题的数值方法

这一节，我们假定 LS 问题(1.1)中的矩阵 A 是亏秩的，即 $\mathrm{rank}(A) < n$。此时 LS 问题(1.1)有无穷多个解，而且上节所介绍的处理满秩 LS 问题的数值方法在这里都是行不通的。因此，这一节我们专门来考虑怎样计算亏秩 LS 问题的解。

3.1 列主元QR分解法

设 LS 问题(1.1)中的已知向量 b 分解为

$$b = b_1 + b_2, \tag{3.1}$$

其中 $b_1 \in \mathscr{R}(A)$，$b_2 \in \mathscr{R}(A)^{\perp}$。则易证问题(1.1)等价于求解

$$Ax = b_1. \tag{3.2}$$

现假定 $\mathscr{R}(A) = \mathscr{R}(Q_1)$，其中 $Q_1 \in \mathbb{R}^{m \times r}$，$Q_1^{\mathrm{T}} Q_1 = I$，即 Q_1 的列构成空间 $\mathscr{R}(A)$ 的一组标准正交基。则存在矩阵 $S \in \mathbb{R}^{r \times n}$ 和向量 $c \in \mathbb{R}^r$，使得

$$A = Q_1 S, \quad b_1 = Q_1 c. \tag{3.3}$$

将(3.3)代入(3.2)，并注意到 Q_1 的列线性无关，即知(3.2)等价

于

$$Sx = c。\tag{3.4}$$

显然，这一方程组总是相容的。此外，从(3.3)可得

$$c = Q_1^T b_1 = Q_1^T(b - b_2) = Q_1^T b,\tag{3.5}$$

$$S = Q_1^T A。\tag{3.6}$$

因此，只要求得$\mathscr{R}(A)$的一组标准正交基，就可通过(3.6)，(3.5)和(3.4)求得 LS 问题(1.1)的任一解。

其实，上节所介绍的正交化方法，实质上也是在求$\mathscr{R}(A)$的正交基，这是因为在 $\mathrm{rank}(A) = n$ 的条件下，分解式(2.4)实际上亦给出了$\mathscr{R}(A)$的一组标准正交基。然而，在 $\mathrm{rank}(A) < n$ 时，分解式(2.4)一般并不一定能够产生$\mathscr{R}(A)$的一组标准正交基。例如，设

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix},$$

则有 $\mathrm{rank}(A) = 2$，而且有分解 $A = I_4 A$。但单位向量 e_1, e_2, e_3, e_4 中的任意两个均非$\mathscr{R}(A)$的标准正交基。

但是，如果我们先对 A 的列进行适当的排列使其前 r 列线性无关，然后再进行 QR 分解，则照样可以产生$\mathscr{R}(A)$的标准正交基。事实上，若有分解

$$AP = Q\begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix}\begin{matrix} r \\ m-r \end{matrix},\tag{3.7}$$
$$\qquad\quad r \qquad n-r$$

其中 P 是排列方阵，Q 是正交矩阵，R_{11} 是非奇异的上三角阵，则 Q 的前 r 列就是$\mathscr{R}(A)$的一组标准正交基。

一旦分解式(3.7)已经求出，则由(3.5)和(3.6)可知，此时有

$$S=[R_{11},R_{12}]P^{\mathrm{T}}, \qquad Q^{\mathrm{T}}b=\begin{bmatrix} c \\ d \end{bmatrix}\begin{matrix} r \\ m-r \end{matrix};$$

再将 S 和 c 代入(3.4)，即知 LS 问题(1.1)的通解为

$$x=P\begin{bmatrix} R_{11}^{-1}(c-R_{12}z) \\ z \end{bmatrix}, \qquad z\in\mathbb{R}^{n-r}. \qquad (3.8)$$

通常，我们把上述表达式中 $z=0$ 时对应的解称作 LS 问题 (1.1) 的**基本解**，记作 x_b，即

$$x_b=P\begin{bmatrix} R_{11}^{-1}c \\ 0 \end{bmatrix}. \qquad (3.9)$$

现在，我们再来考虑如何计算分解式(3.7)。类似于分解式(2.4)的计算，这一分解可用 Householder 变换和适当的列交换相结合逐步求得。假设对某一正整数 k，我们已求得 $k-1$ 个 Householder 变换 $H_1,\cdots,H_{k-1}$ 和 $k-1$ 个初等交换矩阵 $P_1,\cdots,P_{k-1}$，使得

$$\begin{aligned} R_{k-1}&=(H_{k-1}\cdots H_1)A(P_1\cdots P_{k-1}) \\ &=\begin{bmatrix} R_{11}^{(k-1)} & R_{12}^{(k-1)} \\ 0 & R_{22}^{(k-1)} \end{bmatrix}\begin{matrix} k-1 \\ m-k+1 \end{matrix}, \\ &\qquad\ \ \begin{matrix} k-1 \qquad & n-k+1 \end{matrix} \end{aligned} \qquad (3.10)$$

其中 $R_{11}^{(k-1)}$ 是非奇异的上三角阵. 现在记

$$R_{22}^{(k-1)}=[v_k^{(k-1)},\cdots,v_n^{(k-1)}],$$

即 $v_i^{(k-1)}$ 表示 $R_{22}^{(k-1)}$ 的第 $i-k+1$ 列。下一步是，先确定指标 p，$k\leqslant p\leqslant n$，满足

$$\|v_p^{(k-1)}\|_2=\max\{\|v_k^{(k-1)}\|_2,\cdots,\|v_n^{(k-1)}\|_2\}; \qquad (3.11)$$

如果 $\|v_p^{(k-1)}\|_2=0$，则计算结束；否则取 P_k 就是第 k 列与第 p 列交换的初等交换矩阵，并确定一个 Householder 变换 $\tilde{H}_k\in\mathbb{R}^{(m-k+1)\times(m-k+1)}$，使得

$$\tilde{H}_k v_p^{(k-1)}=r_{kk}e_1.$$

令 $H_k=\mathrm{diag}(I_{k-1},\tilde{H}_k)$，则有

$$R_k = H_k R_{k-1} P_k = \begin{bmatrix} R_{11}^{(k)} & R_{12}^{(k)} \\ 0 & R_{22}^{(k)} \end{bmatrix} \begin{matrix} k \\ m-k \end{matrix}, \qquad (3.12)$$
$$\qquad\qquad\qquad\qquad\quad k \qquad n-k$$

其中 $R_{11}^{(k)}$ 是 $k \times k$ 的非奇异上三角阵。这样，从 $k=1$ 出发，依次进行 $r=\mathrm{rank}(A)$次，即可求得分解(3.7)。

此外，(3.11)中的范数也不需每步都按 2 范数的定义去计算，这只需注意到下面的事实即可：

$$Ux = \begin{bmatrix} \alpha \\ y \end{bmatrix} \begin{matrix} 1 \\ l \end{matrix} \Longrightarrow \|x\|_2^2 = \alpha^2 + \|y\|_2^2$$

对任意的正交矩阵 U 成立。

综上所述，可得计算分解(3.7)的算法如下：

算法3.1

(1) 输入 $A=[a_{ij}] \in \mathbb{R}^{m \times n}$；

$$p_j := j, \quad \gamma_j := \sum_{i=1}^{m} a_{ij}^2, \quad j=1,2,\cdots,n, \quad k:=1.$$

(2) 确定 l 使 $\gamma_l = \max\limits_{k \leq j \leq n} \gamma_j$.

(3) 如果 $\gamma_l = 0$，则分解结束；否则进行下一步。

(4) 交换 p_l 与 p_k，γ_l 与 γ_k，以及 a_{il} 与 a_{ik}，$i=1,2,\cdots,m$。

(5) 确定一个 Householder 变换 $\tilde{H}_k$，使得

$$\tilde{H}_k \begin{bmatrix} a_{kk} \\ \vdots \\ a_{mk} \end{bmatrix} = \begin{bmatrix} * \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

(6) $A := \mathrm{diag}(I_{k-1}, \tilde{H}_k)A$，$\gamma_j := \gamma_j - a_{kj}^2$，$j=k+1,\cdots,n$。

(7) $k := k+1$，转步(2)。

这算法的运算量是 $2mnr - r^2(m+n) + \dfrac{2}{3}r^3$，其中 $r=\mathrm{rank}(A)$。

分解式(3.7)的R存储在A的上三角部分，初等交换阵由p_j记录下来；如果需要Q，一般以因子形式存储在A的下三角部分；如果只是利用这一分解去求LS问题之解（通常，用来求基本解），则Q不必存储，只需将计算过程每步产生的Householder变换作用在b上即可。

其次，利用上述方法产生的上三角阵

$$R=\begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1r} & \gamma_{1,r+1} & \gamma_{1n} \\ & \ddots & \vdots & \vdots & \vdots \\ & & \gamma_{rr} & \gamma_{r,r+1} & \gamma_{rn} \\ 0 & & & 0 & \end{bmatrix}$$

的对角元素满足

$$\gamma_{kk}^2 \geqslant \sum_{i=k}^{j} \gamma_{ij}^2, \quad k=1,2,\cdots,r, \quad j=k+1,\cdots,n. \qquad (3.13)$$

此外，如果我们希望求出LS问题(1.1)的最小范数解，则还需将分解(3.7)中的R_{12}化为零。这可通过r次Householder变换来完成，即确定r个Householder变换$Z_1,\cdots,Z_r$和一个排列方阵P_2使得

$$[R_{11}, R_{12}]Z_1\cdots Z_rP_2=[\underset{r}{T}, \underset{n-r}{0}],$$

其中T是上三角阵。现令

$$Z=PZ_1\cdots Z_rP_2.$$

则有

$$Q^{\mathrm{T}}AZ=\begin{bmatrix} T & 0 \\ 0 & 0 \end{bmatrix}\begin{matrix} r \\ m-r \end{matrix}.$$
$$\qquad\qquad\quad r \quad\ n-r$$

由此立即可得

$$x_{\mathrm{LS}}=Z\begin{bmatrix} T^{-1}c \\ 0 \end{bmatrix},$$

其中的c是由$Q^{\mathrm{T}}b$的前r个分量构成的r维向量。请读者作为练

习写出求 x_{LS} 的详细算法。

3.2 奇异值分解法

设 A 的奇异值分解为

$$A = U\Sigma V^{T}, \qquad \Sigma = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \begin{matrix} r \\ m-r \end{matrix}, \quad \begin{matrix} r & n-r \end{matrix}$$

其中 $U=[u_1,\cdots,u_m]$和 $V=[v_1,\cdots,v_n]$是正交阵， $\Sigma_r=\mathrm{diag}(\sigma_1,\cdots,\sigma_r)$， $\sigma_1\geqslant\cdots\geqslant\sigma_r>0$. 则由定理 1.2 知，

$$x_{LS} = A^{\dagger}b = V\begin{bmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{bmatrix}U^{T}b$$

$$= \sum_{i=1}^{r}\frac{u_i^{T}b}{\sigma_i}v_i. \qquad (3.14)$$

因此，一旦求出分解(3.13)，我们就可由(3.14)很容易求出LS问题(1.1)的最小范数解 x_{LS}。关于分解(3.13)的具体计算 方 法，我们将在第七章中详细讨论.

3.3 数值秩的定义和确定方法

从上面的讨论可知，亏秩 LS 问题之解的确定与矩阵 的秩密切相关。然而，“秩”这一在数学上精确定义的概念，在有误差出现的计算机上就变得模糊不清了。因此，我们需要引进数值秩的概念。

定义3.1 设 $A\in\mathbb{R}^{m\times n}$. 如果对某一正数 δ 有

$$r=\min\{\mathrm{rank}(B):\ B\in\mathbb{R}^{m\times n},\ \|B-A\|_2\leqslant\delta\}, \qquad (3.15)$$

则称 r 是矩阵 A 的 **δ数值秩.**

设 $A\in\mathbb{R}^{m\times n}$ 的奇异值是 $\sigma_1\geqslant\cdots\geqslant\sigma_n$ （这里假定$m\geqslant n$)，对应的左、右奇异向量分别为 $u_1,\cdots,u_n$ 和 $v_1,\cdots,v_n$. 则由

$$\inf_{\mathrm{rank}(B)<k}\|A-B\|_2=\sigma_{k+1}, \qquad k=1,2,\cdots,n,$$

而且下确界在

$$B=\sum_{i=1}^{k}\sigma_i u_i v_i^{\mathrm{T}}$$

达到，可知矩阵A的δ数值秩是r的充分必要条件是

$$\sigma_1\geqslant\cdots\geqslant\sigma_r>\delta\geqslant\sigma_{r+1}\geqslant\cdots\geqslant\sigma_n. \qquad (3.16)$$

因此，我们可以应用奇异值分解来确定矩阵的δ数值秩。由于奇异值的良好数值性态，目前人们普遍认为奇异值分解法是确定数值秩的最可靠方法。

如果用第七章的算法 5.2 于A上计算的奇异值满足

$$\hat{\sigma}_1\geqslant\cdots\geqslant\hat{\sigma}_{\hat{r}}>\delta\geqslant\hat{\sigma}_{\hat{r}+1}\geqslant\cdots\geqslant\hat{\sigma}_n, \qquad (3.17)$$

则根据误差分析和奇异值的扰动分析的结果我们可以认为A的δ数值秩就是$\hat{r}$，而且亦可用

$$x_{\hat{r}}=\sum_{i=1}^{\hat{r}}\frac{\hat{u}_i^{\mathrm{T}}b}{\sigma_i}\hat{v}_i \qquad (3.18)$$

作为 LS 问题(1.1)之最小范数解x_{LS}的近似值，其中$\hat{u}_i$和$\hat{v}_i$分别是左、右奇异向量u_i和v_i的计算值。

一般δ应该与机器的精度相容，通常可取$\delta=\varepsilon\|A\|_\infty$；但如果数据本身的误差就比$\varepsilon$大得多，那么此时$\delta$也应取得大一些，比如可取$\delta=10^{-2}\|A\|_\infty$；此外，由于

$$\|x_{\hat{r}}\|_2\approx\frac{1}{\hat{\sigma}_{\hat{r}}}\leqslant\frac{1}{\delta},$$

故根据实际需要δ亦可选得使$\|x_{\hat{r}}\|_2$适当的小。

这里需要指出的一点是，欲使(3.16)成立，σ_{r+1}与σ_r之间必须具有一定的距离才行。因此，当A的奇异值分离不明显，而又是亏秩矩阵时，应用奇异值分解法来确定矩阵A的δ数值秩就会遇到一定的困难，就需要更复杂的方法来处理(参阅文献[68])。

现在我们来考虑算法 3.1，在误差出现的情况下，如何较合理地终止分解。此时，即使A的准确秩是k，一般进行k步之

后，$R_{22}^{(k)}$ 也不会等于零。但如果 $R_{22}^{(k)}$ 相对于A很小的话，比如对某一较小的正数 ε，有

$$\|R_{22}^{(k)}\|_2 \leqslant \varepsilon \|A\|_2 = \varepsilon\sigma_1, \tag{3.19}$$

则我们应该有理由认为A的数值秩是k。事实上，如果(3.19)满足，那么，根据第一章的推论 6.6，可知

$$\sigma_{k+1}(R_k) \leqslant \|R_{22}^{(k)}\|_2 \leqslant \varepsilon\sigma_1, \tag{3.20}$$

其中的 $\sigma_{k+1}(R_k)$表示(3.12)所定义的矩阵 R_k 的第 $k+1$ 个最大奇异值。再利用关于 Householder 变换的误差分析结果，易证计算得到的 R_k 满足

$$QR_k = A + E_k,$$

其中Q是正交矩阵，E_k 满足

$$\|E_k\|_2 \leqslant c_1 \varepsilon n^{1/2} \sigma_1,$$

此处 c_1 是与 m, n 有关的常数（参见文献[20]）。于是

$$\begin{aligned} \sigma_{k+1} &\leqslant \sigma_{k+1}(R_k) + c_1 u n^{1/2} \sigma_1 \\ &\leqslant (\varepsilon + c_1 \varepsilon n^{1/2}) \sigma_1. \end{aligned}$$

这表明，矩阵A关于正数 $\delta = (\varepsilon + c_1 \varepsilon n^{1/2})\sigma_1$ 的数值秩至多是k。

另一方面，从不等式(3.13)可得

$$\|R_{22}^{(k)}\|_2 \leqslant \|R_{22}^{(k)}\|_F \leqslant (n-k)^{1/2} |\gamma_{k+1,k+1}|,$$

$k = 1, 2, \cdots, n$。而

$$|\gamma_{11}| \leqslant \sigma_1,$$

因此，如果对某一正数 ε_0 有

$$|\gamma_{k+1,k+1}| \leqslant \varepsilon_0 |\gamma_{11}|, \tag{3.21}$$

则对正数 $\varepsilon = (n-k)^{\frac{1}{2}} \varepsilon_0$ 有(3.19)成立。因而通常以(3.21)作为算法 3.1 实际使用时的停机准则。

上述分析表明，如果列主元 QR 分解中的上三角矩阵R的第 $k+1$个对角元素很小，则A与一个秩为k的矩阵很靠近。然而上述结论反过来不真，即一个与秩为k的矩阵很靠近的矩阵，其列

主元 QR 分解中 R 的第 $k+1$ 个对角元素不一定很小，请看下例。

例3.1 (Kahan，参见文献[45]) 设

$$R_n = \mathrm{diag}(1, s, \cdots, s^{n-1}) \begin{bmatrix} 1 & -c & -c & \cdots & -c \\ & 1 & -c & \cdots & -c \\ & & \ddots & \ddots & \vdots \\ & 0 & & \ddots & -c \\ & & & & 1 \end{bmatrix},$$

其中 $s^2+c^2=1$，$s,c>0$。对于 $n=100$，$c=0.2$，有 R_n 的最小奇异值为 $\sigma_n \approx 0.368\times 10^{-8}$，即 R_n 已与奇异矩阵很接近，但 $\gamma_{nn}=s^{n-1}\approx 0.133$。

由此可见，列主元 QR 分解作为一种判定矩阵的数值秩的方法从理论上来讲是不可靠的。但是，从实际计算的经验来看，它一般都能得到令人满意的结果。

最后，我们顺便指出的一点是，列主元 QR 分解亦可用来估计矩阵的条件数。设 $A\in\mathbb{R}^{n\times n}$ 有分解：

$$QAP=R,$$

其中 Q 为正交矩阵，P 为排列方阵，R 是上三角矩阵，且其对角元素满足

$$\gamma_{11}\geqslant\gamma_{22}\geqslant\cdots\geqslant\gamma_{nn}>0,$$

则易证

$$\sigma_n(A)=\sigma_n(R)\leqslant\gamma_{nn},$$
$$\sigma_1(A)=\sigma_1(R)\geqslant\gamma_{11}.$$

所以，有

$$\kappa_2(A)=\frac{\sigma_1(A)}{\sigma_n(A)}\geqslant\frac{\gamma_{11}}{\gamma_{nn}}.$$

因此，我们可用 γ_{11}/γ_{nn} 作为谱条件数 $\kappa_2(A)$ 的估计。Stewart(参见文献[63])所作的数值实验的结果表明：这一估计对于揭示 $\kappa_2(A)$ 的数量级是十分可靠的，一般只比真实条件数少 2 到 3 倍，而

不会超过10倍。

§4 求解LS问题的迭代法

对于某些大型稀疏的 LS 问题，与线性方程组的情形一样，用迭代法来求解是非常有效的。因此，这一节我们就从两个方面来介绍求解大型稀疏 LS 问题(1.1)的迭代法。这里假定 $A\in\mathbb{R}_n^{m\times n}$，即 $\mathrm{rank}(A)=n$。

4.1 基于正规化方程组的古典迭代法

原则上讲，任何实用于对称正定线性方程组的迭代方法都可应用于 LS 问题(1.1)的正规化方程组

$$A^{\mathrm{T}}Ax=A^{\mathrm{T}}b, \tag{4.1}$$

而得到相应的求解 LS问题(1.1)的迭代法。下面就将第四章所介绍的古典迭代法应用于(4.1)所得到的迭代公式简述如下：

为了叙述简洁，我们记 A 的第 j 列是 a_j，即

$$A=[a_1,a_2,\cdots,a_n],$$

并记

$$d_j=a_j^{\mathrm{T}}a_j,\qquad j=1,2,\cdots,n, \tag{4.2}$$

$$D=\mathrm{diag}(d_1,\cdots,d_n). \tag{4.3}$$

将 Jacobi 迭代法应用于(4.1)，并适当整理，可得到求解LS问题(1.1)的 Jacobi 迭代法的基本公式为：

$$x_{k+1}=x_k+D^{-1}A^{\mathrm{T}}(b-Ax_k),\qquad k=0,1,2,\cdots, \tag{4.4}$$

其中 x_0 为初始向量。

相应的 Gauss-Seidel 迭代法的基本迭代公式为：

$$\begin{aligned}
&z_1=x_k,\ \ r_1=b-Ax_k,\\
&\left.\begin{aligned}
&\delta_j=a_j^{\mathrm{T}}r_j/d_j,\\
&z_{j+1}=z_j+\delta_je_j,\\
&r_{j+1}=r_j-\delta_ja_j,
\end{aligned}\right\}\quad j=1,2,\cdots,n, \qquad\qquad (4.5)\\
&x_{k+1}:=z_{n+1};\qquad\qquad k=0,1,2,\cdots.
\end{aligned}$$

相应的 SOR 迭代法，只需在上述迭代公式中将 δ_j 换为

$$\delta_j = \omega a_j^{\mathrm{T}} r_j / d_j, \quad 0 < \omega < 2 \tag{4.6}$$

即可，其中 ω 是松弛参数。

此外，第五章所讲的共轭梯度法可以逐字不变地搬过来，这里不再赘述。当然，为了加速收敛应与各种预优技巧结合起来使用。

4.2 基于等价方程组的 SOR 和 SSOR 迭代法

由于(4.1)亦可写成如下等价形式，求一个 n 维向量 x 和一个 m 维向量 r，使得

$$\begin{cases} r + Ax = b, \\ A^{\mathrm{T}} r = 0. \end{cases} \tag{4.7}$$

因此，我们亦可利用(4.7)来构造求解 LS 问题(1.1)的迭代法。这里，将着重考虑基于(4.7)的 SOR 和 SSOR 迭代法。

现不妨假定 A 具有如下形状：

$$A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \begin{matrix} n \\ m-n \end{matrix}, \tag{4.8}$$

其中 A_1 是 $n \times n$ 非奇异矩阵（如若不然，可适当调整(4.7)中方程的次序而使(4.8)满足）。再把 r 和 b 作相应的分块

$$r = \begin{bmatrix} v \\ w \end{bmatrix}, \qquad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix},$$

其中 $v, b_1 \in \mathbb{R}^n$，$w, b_2 \in \mathbb{R}^{m-n}$。则(4.7)可写成

$$Cz = d, \tag{4.9}$$

其中

$$C = \begin{bmatrix} A_1 & 0 & I \\ A_2 & I & 0 \\ 0 & A_2^{\mathrm{T}} & A_1^{\mathrm{T}} \end{bmatrix} \begin{matrix} n \\ m-n \\ n \end{matrix}, \quad z = \begin{bmatrix} x \\ w \\ v \end{bmatrix}, \quad d = \begin{bmatrix} b_1 \\ b_2 \\ 0 \end{bmatrix}.$$

$$\qquad\quad n \qquad m-n \qquad n$$

由于 A_1 是非奇异的，因而不难验证 C 亦是非奇异的。

这里顺便指出，方程组(4.9)常常用来迭代改进 LS问题的近

似解，而且一般效果都很好。假定已知 LS问题(1.1)的一个近似解 $\hat{x}$，记 $\hat{r}=b-A\hat{x}$，$\hat{x}$ 和 $\hat{r}$ 对应的向量 z 为 $\hat{z}$。用双精度计算

$$c=d-C\hat{z},$$

再用单精度求解方程组

$$C\tilde{z}=c,$$

则 $\hat{x}+\tilde{x}$ 为 $\hat{x}$ 的改进，其中 $\tilde{x}$ 为 $\tilde{z}$ 的前 n 个分量构成的 n 维向量。上述过程可以重复，直到改进的近似解满足要求的精度。

现在，再回到我们的主题：考虑基于(4.9)的分块 SOR 和 SSOR 迭代法。

按照 C 的自然分块，对应于第四章中的记号，有

$$D=\operatorname{diag}(A_1,I,A_1^{\mathrm{T}}),\tag{4.10a}$$

$$C_L=-\begin{bmatrix}0&0&0\\A_2&0&0\\0&A_2^{\mathrm{T}}&0\end{bmatrix},\quad C_U=-\begin{bmatrix}0&0&I\\0&0&0\\0&0&0\end{bmatrix}\tag{4.10b}$$

$$L=D^{-1}C_L=\begin{bmatrix}0&0&0\\B_2&0&0\\0&B_3&0\end{bmatrix},\quad U=D^{-1}C_U=\begin{bmatrix}0&0&B_1\\0&0&0\\0&0&0\end{bmatrix},\tag{4.10c}$$

其中 $B_1=-A_1^{-1}$，$B_2=-A_2$，$B_3=-(A_2A_1^{-1})^{\mathrm{T}}$。

对应的 Jacobi 迭代矩阵为：

$$J=L+U。\tag{4.11}$$

直接计算有

$$\begin{aligned}J^3&=\operatorname{diag}(B_1B_3B_2,B_2B_1B_3,B_3B_2B_1)\\&=-\operatorname{diag}(A_1^{-1}B^{\mathrm{T}}BA_1,BB^{\mathrm{T}},B^{\mathrm{T}}B),\end{aligned}\tag{4.12}$$

其中 $B=A_2A_1^{-1}$。

对应的 SOR 迭代矩阵为：

$$\mathscr{L}_\omega=(I-\omega L)^{-1}[(1-\omega)I+\omega U]。\tag{4.13}$$

将(4.10c)代入(4.13)有

$$\mathscr{L}_\omega = \begin{bmatrix} I & 0 & 0 \\ -\omega B_2 & I & 0 \\ 0 & -\omega B_3 & I \end{bmatrix}^{-1} \begin{bmatrix} (1-\omega)I & 0 & \omega B_1 \\ 0 & (1-\omega)I & 0 \\ 0 & 0 & (1-\omega)I \end{bmatrix}$$

$$= (1-\omega)I + K_\omega, \tag{4.14}$$

其中

$$K_\omega = \begin{bmatrix} 0 & 0 & \omega B_1 \\ \omega(1-\omega)B_2 & 0 & \omega^2 B_2 B_1 \\ (1-\omega)\omega^2 B_3 B_2 & \omega(1-\omega)B_3 & \omega^3 B_3 B_2 B_1 \end{bmatrix}.$$

对应的 SSOR 迭代矩阵为：

$$\begin{aligned} \mathscr{S}_\omega &= (I-\omega U)^{-1}[(1-\omega)I+\omega L](I-\omega L)^{-1}[(1-\omega)I+\omega U] \\ &= (I-\omega U)^{-1}(I-\omega L)^{-1}[(1-\omega)I+\omega L][(1-\omega)I+\omega U]. \end{aligned} \tag{4.15}$$

因此，$\mathscr{S}_\omega$ 相似于矩阵

$$\begin{aligned} \hat{\mathscr{S}}_\omega &= (I-\omega L)^{-1}[(1-\omega)I+\omega L][(1-\omega)I+\omega U][I-\omega U]^{-1} \\ &= (I-\omega L)^{-1}[(1-\omega)I+\omega L](I-\omega U)^{-1}[(1-\omega)I+\omega U] \\ &= M(L)M(U), \end{aligned} \tag{4.16}$$

其中

$$M(Z) = (I-\omega Z)^{-1}((1-\omega)I+\omega Z). \tag{4.17}$$

直接计算可得

$$M(L) = \begin{bmatrix} (1-\omega)I & 0 & 0 \\ \omega(2-\omega)B_2 & (1-\omega)I & 0 \\ \omega^2(2-\omega)B_3B_2 & \omega(2-\omega)B_3 & (1-\omega)I \end{bmatrix},$$

$$M(U) = \begin{bmatrix} (1-\omega)I & 0 & \omega(2-\omega)B_1 \\ 0 & (1-\omega)I & 0 \\ 0 & 0 & (1-\omega)I \end{bmatrix}.$$

从而有

$$\hat{\mathscr{S}}_\omega = M(L)M(U) = (1-\omega)^2 I + T_\omega, \tag{4.18}$$

其中

$$T_\omega = \begin{bmatrix} 0 & 0 & \sigma B_1 \\ \sigma B_2 & 0 & \omega^2(2-\omega^2)B_2B_1 \\ \omega\sigma B_3B_2 & \sigma B_3 & \omega^2(2-\omega)^2B_3B_2B_1 \end{bmatrix},$$

此处
$$\sigma = \omega(\omega-1)(\omega-2).$$

定理4.1 设 $J, \mathscr{L}_\omega$ 和 $\mathscr{S}_\omega$ 分别由(4.11)，(4.14)和(4.15)给出。则有：

(1) $\lambda(J^3)\subset[-\rho(J)^3, 0]$，且 J^3 与 $B_3B_2B_1$ 有相同的非零特征值。

(2) 若 $\lambda\in\lambda(\mathscr{L}_\omega)$，$\omega\neq 0, \lambda\neq 0$,则存在 $\mu\in\lambda(J)$使 λ 是方程

$$(\lambda+\omega-1)^3 = \lambda^2\omega^3\mu^3 \tag{4.19}$$

的非零解；反之，对任意的 $\mu\in\lambda(J)$，若 λ 满足(4.19)，则必有 $\lambda\in\lambda(\mathscr{L}_\omega)$。

(3) 若 $\lambda\in\lambda(\mathscr{S}_\omega)$，$\omega\neq 0$，$\lambda\neq 0$，则存在 $\mu\in\lambda(J)$使得 λ 是方程

$$\omega^3(2-\omega)^2\lambda(\lambda+1-\omega)\mu^3 = [\lambda-(1-\omega)^2]^3 \tag{4.20}$$

的非零解；反之，对任意的 $\mu\in\lambda(J)$，若 λ 满足(4.20)，则必有 $\lambda\in\lambda(\mathscr{L}_\omega)$。

证明 (1) 由(4.12)知 $-J^3$ 相似于一个半正定矩阵，从而有

$$\lambda(J^3)\subset[-\rho(J)^3, 0].$$

又 $\lambda(J^3) = \lambda(B_1B_3B_2)\cup\lambda(B_2B_1B_3)\cup\lambda(B_3B_2B_1)$，而 $B_1B_3B_2$，$B_2B_1B_3$ 和 $B_3B_2B_1$ 有相同的非零特征值，因此 J^3 与 $B_3B_2B_1$ 有相同的非零特征值，即(1)得证。

(2) 设 $\lambda\in\lambda(\mathscr{L}_\omega)$，$\lambda\neq 0$，$\omega\neq 0$，并假定对应的特征向量为 $u=(x^T, y^T, z^T)^T$，其中 $x, z\in\mathbb{R}^n, y\in\mathbb{R}^{m-n}$，即

$$\mathscr{L}_\omega u = \lambda u, \qquad u\neq 0. \tag{4.21}$$

将(4.14)代入(4.21)即有

$$(1-\omega)u + K_\omega u = \lambda u,$$

即有

$$K_\omega u=(\lambda+\omega-1)u.$$

记 $\tau=\lambda+\omega-1$，注意到 K_ω 的定义，上式即为

$$\begin{cases}\omega B_1 z=\tau x, & (4.22\text{a})\\ \omega(1-\omega)B_2x+\omega^2B_2B_1z=\tau y, & (4.22\text{b})\\ (1-\omega)\omega^2B_3B_2x+\omega(1-\omega)B_3y & \\ \quad +\omega^3B_3B_2B_1z=\tau z. & (4.22\text{c})\end{cases}$$

现假定 $\tau\neq0$. 由(4.22a)可得 $x=\tau^{-1}\omega B_1z$，代入(4.22b)，得 $y=\tau^{-2}(\omega^2(1-\omega)+\tau\omega^2)B_2B_1z$；再将 x 和 y 代入(4.22c)，即得 $\{\tau^{-1}\omega^3(1-\omega)+\tau^{-2}[\omega^2(1-\omega)+\tau\omega^2]\omega(1-\omega)+\omega^3\}B_3B_2B_1z=\tau z$.
上式两边乘以 τ^2，并将 $\tau=\lambda+\omega-1$ 代入，再整理，可得

$$\omega^3\lambda^2B_3B_2B_1z=(\lambda+\omega-1)^3z. \tag{4.23}$$

注意到 $z\neq0$（否则，由(4.22)可推出亦有 $x=0$，$y=0$，这与 $u\neq0$ 的假定矛盾），(4.23)就表示

$$\frac{(\lambda+\omega-1)^3}{\omega^3\lambda^2}\in\lambda(B_3B_2B_1).$$

再由定理的(1)成立即知，必存在 $\mu\in\lambda(J)$，使得

$$\mu^3=\frac{(\lambda+\omega-1)^3}{\omega^3\lambda^2},$$

即 λ 是方程(4.19)的非零解.

当 $\tau=0$ 时，即 $\lambda=1-\omega$ 时，由 $\lambda\neq0$，故 $\omega\neq1$. 于是由(4.22)知，必有

$$B_1z=0,\qquad B_2x=0,\qquad B_3y=0,$$

即有

$$Ju=0.$$

而 $u\neq0$，故 $0\in\lambda(J)$. 因此，此时取 $\mu=0$ 即有 λ 满足(4.19).

反之，对任意给定的 $\mu\in\lambda(J)$，设 λ 满足(4.19). 当 $\lambda\neq0$ 且 $\omega\neq0$ 时，将上述证明过程倒推回去即知 $\lambda\in\lambda(\mathscr{L}_\omega)$；当 $\lambda\neq0$ 而

$\omega=0$ 时，有 $\lambda=1$，$\mathscr{L}_0=I$，当然亦有 $\lambda\in\lambda(\mathscr{L}_\omega)$；当 $\lambda=0$ 时，必有 $\omega=1$，又 $\mathscr{L}_1$ 是奇异矩阵，所以 $\lambda=0\in\lambda(\mathscr{L}_1)$。这样(2)得证。

至于(3)，只要注意到 $\mathscr{S}_\omega$ 和 $\hat{\mathscr{S}}_\omega$ 相似，而且 $\hat{\mathscr{S}}_\omega$ 有表达式(4.18)，就可完全类似于上述推理过程证之，详细证明留作练习。证毕。

为了给出 SOR 和 SSOR 迭代法的收敛性定理，我们引用一个经典的结果：

引理4.1 实系数三次多项式

$$t^3+pt^2+qt+r$$

的三个根的绝对值都小于1的充分必要条件是：

(1) $1+r>0$，$1-r>0$；

(2) $1+p+q+r>0$，$1-p+q-r>0$；

(3) $1-q+pr-r^2>0$。

证明参见文献[7]。

令 $\beta=\rho(J)$ 表示(4.11)所定义的分块 Jacobi迭代矩阵 J 的谱半径。对分块 SOR 迭代我们有：

定理4.2 用分块 SOR 迭代法求解(4.9)时，当且仅当 (β,ω) 位于如图 4.1 所示的区域内部时迭代法收敛。

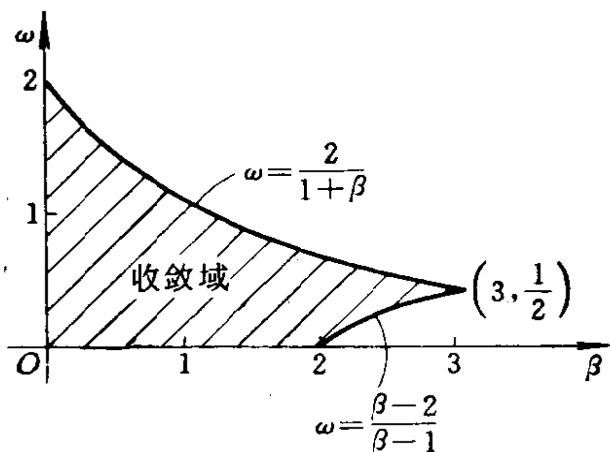


图 4.1

注4.1 (1) 从图 4.1 不难看出，当分块 Jacobi 迭代法发散时（即 $1\leqslant\beta\leqslant3$ 时），分块 SOR 迭代法依然可能收敛。

(2) Varga 等(1984)已证明了：将分块 SOR迭代法应用于方程组(4.9)时的最佳松弛因子 ω_b(即 $\rho(\mathscr{L}_{\omega_b})=\min\limits_{0<\omega<2}\rho(\mathscr{L}_\omega)$) 是下面关于 ω 的三次方程

$$4\beta^3\omega^3+27\omega-27=0,\qquad 0\leqslant\beta<3 \tag{4.24a}$$

的唯一正根，而且满足

$$\frac{1}{2}<\omega_b\leqslant1, \tag{4.24b}$$

$$\rho(\mathscr{L}_{\omega_b})=2(1-\omega_b). \tag{4.24c}$$

详见文献[52]。

关于分块 SSOR 迭代法，有如下收敛性定理：

定理4.3 用分块 SSOR 迭代法求解(4.9)时，当且仅当(β,ω)位于如图 4.2 所示的区域内部时迭代法收敛。

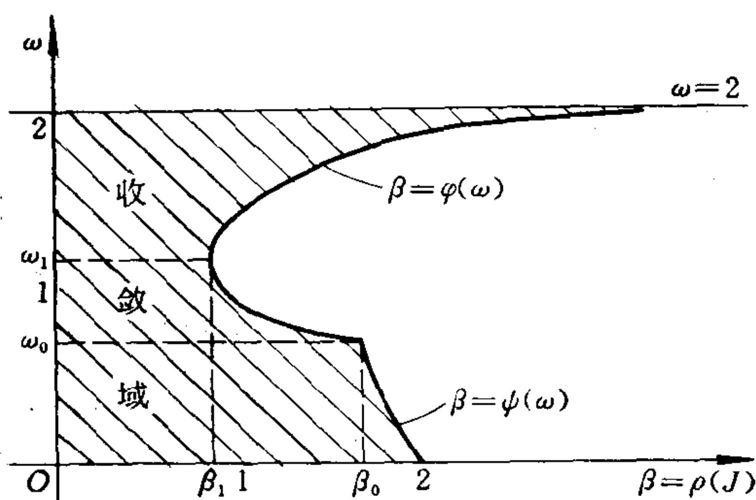


图 4.2

这里需要说明的是，在图 4.2 中

$$\varphi(\omega)=[1+(1-\omega)^2]/\omega[\omega(2-\omega)^2]^{1/3},\quad 0<\omega<2, \tag{4.25a}$$

$$\psi(\omega)=[1+(1-\omega)^2]\{(2-\omega)/(1-\omega)[1+(1-\omega)^5]\}^{1/3},$$

$$0<\omega<1, \tag{4.25b}$$

$$\omega_0=\frac{1}{2}[\sqrt{5}-1],\ \omega_1=4-2\sqrt{2}, \tag{4.25c}$$

$$\beta_0=\varphi(\omega_0)=\psi(\omega_0)=3/5^{1/3}, \tag{4.25d}$$

$$\beta_1=\varphi(\omega_1)<\varphi(1)=1. \tag{4.25e}$$

注4.2 从图 4.2 可以看出，无论 $\beta>0$ 为何值，总可选择一个参数 ω 使得 SSOR 迭代法收敛；由此可见，当 SOR 迭代法不收敛时，亦可构造出收敛的 SSOR 迭代法.

下面我们只给出定理 4.3 的证明，有关定理 4.2 的证明可以类似给出，详细证明留作练习.

定理4.3的证明 由熟知的迭代收敛准则知，用分块SSOR迭代法求解方程组(4.9)时，收敛的充分必要条件是迭代矩阵的谱半径小于 1，即

$$\rho(\mathscr{S}_\omega)<1,$$

其中 $\mathscr{S}_\omega$ 由(4.15)给出. 再应用定理4.1 的(3)知，对任意的$\omega\in(0,2)$，有

$$\rho(\mathscr{S}_\omega)=\max_{\mu\in\lambda(J)}\{|\lambda|:\ \omega^3(2-\omega)^2\lambda(\lambda+1-\omega)\mu^3=[\lambda-(1-\omega)^2]^3\}. \tag{4.26}$$

现令

$$\begin{aligned}f(\lambda)&=[\lambda-(1-\omega)^2]^3+\lambda(\lambda+1-\omega)\omega^3(2-\omega)^2\mu^3\\&=\lambda^3+p\lambda^2+q\lambda+r,\end{aligned} \tag{4.27}$$

其中

$$p=(2-\omega)^2\omega^3\mu^3-3(1-\omega)^2,$$

$$q=(1-\omega)(2-\omega)^2\omega^3\mu^3+3(1-\omega)^4,$$

$$r=-(1-\omega)^6.$$

注意到 $\lambda(J^3)\subset[-\rho(J)^3,0]$，根据引理 4.1，由(4.26)和(4.27)知，$\rho(\mathscr{S}_\omega)<1$ 的充分必要条件是，对任意的 $\mu\in\lambda(-J)\subset[0,\rho(J)]$有

$$
\begin{cases}
(1-\omega)^6<1, \\
[1-(1-\omega)^2]^3+\omega^3\mu^3(2-\omega)^3>0, \\
[1+(1-\omega)^2]^3-\omega^4\mu^3(2-\omega)^2>0, \\
[1-(1-\omega)^4]^3-(2-\omega)^2\omega^3\mu^3[1-\omega)+(1-\omega)^6]>0.
\end{cases}
\tag{4.28}
$$

通过不太复杂的初等运算，可证不等式组(4.28)对任意的 $\mu \in \lambda(-J)$ 成立的充分必要条件是：

(i) 当 $0<\omega<1$ 时，$0\leqslant\beta<\min\{\varphi(\omega),\psi(\omega)\}$； (4.29a)

(ii) 当 $1\leqslant\omega<2$ 时，$0\leqslant\beta<\varphi(\omega)$， (4.29b)

其中 φ 和 ψ 由(4.25)给出，$\beta=\rho(J)$。

对 $\varphi(\omega)$ 微分得

$$\varphi'(\omega)=f(\omega)g(\omega),$$

其中

$$g(\omega)=2[(1-\omega)^2+1]^2/[3\omega^5(2-\omega)^3\varphi^2(\omega)],$$

$$f(\omega)=-(\omega-4)^2+8.$$

而 $g(\omega)$ 在 $0<\omega<2$ 之内恒大于零，因此立即有，$\varphi(\omega)$ 在 $0<\omega<2$ 之内有唯一的极小点 $\omega_1=4-2\sqrt{2}$，且在 $0<\omega<\omega_1$ 内 $\varphi(\omega)$ 严格下降，在 $\omega_1<\omega<2$ 内 $\varphi(\omega)$ 严格上升，并以 $\omega=2$ 为渐近线。

对 $\psi(\omega)$ 微分得

$$\psi'(\omega)=\hat{f}(\omega)\hat{g}(\omega),$$

其中

$$\hat{g}(\omega)=\frac{(2-\omega)^3[(1-\omega)^2+1]^2[1-(1-\omega)^3]}{3[(1-\omega)+(1-\omega)^6]^2\psi^2(\omega)},$$

$$\hat{f}(\omega)=\omega^2+\omega-1.$$

而 $\hat{g}(\omega)$ 在 $0<\omega<1$ 之内恒大于零，故 $\psi(\omega)$ 在 $0<\omega<1$ 之内亦有唯一的极小点 $\omega_0=\dfrac{1}{2}(\sqrt{5}-1)$，而且在 $0<\omega<\omega_0$ 之内 $\psi(\omega)$ 严格下降，在 $\omega_0<\omega<1$ 之内 $\psi(\omega)$ 严格上升。

再令 $\varphi(\omega)=\psi(\omega)$，注意到 $0<\omega<1$，适当整理可得

$$\omega^4(2-\omega)^3=(1-\omega)+(1-\omega)^6.$$

解此方程可知，在 $0<\omega<1$ 内有唯一的解

$$\omega_0=\frac{1}{2}(\sqrt{5}-1),$$

且易算出 $\beta_0=\psi(\omega_0)=\varphi(\omega_0)=3/5^{1/3}$.

据 $\psi(\omega)$和$\varphi(\omega)$的上述性质和(4.29)，我们可以画出如图4.2所示的收敛区域。

§5　完全最小二乘问题

所谓**完全最小二乘问题**(简称 TLS 问题）是指如下的优化问题：求 $E\in\mathbb{R}^{m\times n}$ 和 $r\in\mathbb{R}^m$，使得

$$\|[E,r]\|_F=\min, \tag{5.1a}$$

$$\text{s. t.}\quad b+r\in\mathscr{R}(A+E), \tag{5.1b}$$

其中 $A\in\mathbb{R}^{m\times n}$ 和 $b\in\mathbb{R}^m$ 已知。

如果 TLS 问题(5.1)有解$[E,r]$，那么，我们称满足

$$(A+E)x=b+r$$

的 x 为 TLS 问题(5.1)的**完全最小二乘解**（简称 TLS 解）。

在统计计算，回归分析和非线性最优方法中都会遇到上述的TLS问题。

前面几节所讨论的 LS 问题(1.1)亦可等价地表述为：求 $r\in\mathbb{R}^m$，使得

$$\|r\|_2=\min, \tag{5.2a}$$

$$\text{s. t.}\quad b+r\in\mathscr{R}(A). \tag{5.2b}$$

事实上，有 $x\in\mathscr{X}_{LS}$当且仅当$Ax=b+r$，其中 r 为优化问题(5.2)之解。因此，TLS 问题 (5.1) 可以看作是 LS 问题 (1.1) 的自然推广。但是，与 LS 问题(1.1)不同，并非对所有给定的数据 A 和 b，TLS 问题(5.1)均有解。例如，设

$$A=\begin{bmatrix}1&0\\0&0\end{bmatrix},\qquad b=\begin{bmatrix}1\\1\end{bmatrix}.$$

则对任意的 $\varepsilon>0$，都有 $b\in\mathscr{R}(A+E_\varepsilon)$，其中 $E_\varepsilon=\mathrm{diag}(0,\varepsilon)$，但 $b\notin\mathscr{R}(A)$。因此，不存在 $E\in\mathbb{R}^{2\times 2}$ 和 $r\in\mathbb{R}^2$ 使(5.1)成立，即 TLS 问题此时无解。

这样，我们首先需要解决的问题就是 TLS 问题(5.1)何时有解。为此，我们先证几个引理。

引理5.1 设 $B\in\mathbb{R}^{n\times n}$ 是对称正定矩阵，$x\in\mathbb{R}^n$ 满足 $\|x\|_2=1$。如果 $\lambda=x^{\mathrm{T}}Bx$ 是 B 的最小(或最大)特征值，则 x 是属于 λ 的单位特征向量。

证明留作练习。

引理5.2 设 $B\in\mathbb{R}^{m\times n}$ 的最小奇异值是 σ，$E\in\mathbb{R}^{m\times n}$ 是满足条件 $\|E\|_F=\sigma$ 的任一矩阵。如果存在非零向量 $v\in\mathbb{R}^n$ 使得 $(A+E)v=0$，则 v 必是 B 之属于 σ 的右奇异向量。

证明 不妨假定 $\|v\|_2=1$。根据第一章的习题 7，可得

$$\sigma=\min_{\|x\|_2=1}\|Bx\|_2\leqslant\|Bv\|_2=\|-Ev\|_2\leqslant\|E\|_2\leqslant\|E\|_F=\sigma.$$

因此，

$$\|Bv\|_2=\sigma. \tag{5.3}$$

如果 $\sigma=0$，则(5.3)蕴含着 $Bv=0$，从而 v 是属于 σ 的右奇异向量；如果 $\sigma\neq 0$，则(5.3)蕴含着

$$\sigma^2=v^{\mathrm{T}}B^{\mathrm{T}}Bv,$$

由引理 5.1 即知，v 是 $B^{\mathrm{T}}B$ 之属于 σ^2 的特征向量，即 v 是 B 之属于 σ 的右奇异向量。证毕。

引理5.3 设 $A\in\mathbb{R}^{m\times n}(m>n)$，$b\in\mathbb{R}^m$，并假定 σ_{n+1} 是矩阵 $[A,b]$ 的最小奇异值。则

$$\inf_{b+r\in\mathscr{R}(A+E)}\|[E,r]\|_F=\sigma_{n+1}. \tag{5.4}$$

证明 记：

$$d=\inf_{b+r\in\mathscr{R}(A+E)}\|[E,r]\|_F. \tag{5.5}$$

由第一章习题 4 知，

$$\sigma_{n+1}=\min_{\mathrm{rank}([A+E,b+r])\leqslant n}\|[E,r]\|_F,$$

而 $b+r\in\mathscr{R}(A+E)$ 蕴含着 $\text{rank}([A+E,b+r])\leqslant n$，从而有

$$d\geqslant\sigma_{n+1}. \tag{5.6}$$

现假定 $\tilde{E}\in\mathbb{R}^{m\times n}$，$\tilde{r}\in\mathbb{R}^m$ 满足：

$$\|[\tilde{E},\tilde{r}]\|_F=\sigma_{n+1} \text{ 且 } \text{rank}([A+\tilde{E},b+\tilde{r}])\leqslant n.$$

则必存在 $x\in\mathbb{R}^n$ 和 $\alpha\in\mathbb{R}$ 不全为零使得

$$(A+\tilde{E})x+\alpha(b+\tilde{r})=0. \tag{5.7}$$

若 $\alpha\neq0$，则(5.7)蕴含着 $b+\tilde{r}\in\mathscr{R}(A+\tilde{E})$，从而

$$d\leqslant\|[\tilde{E},\tilde{r}]\|_F=\sigma_{n+1},$$

结合(5.6)即知 $d=\sigma_{n+1}$，即引理成立；若 $\alpha=0$，则从(5.7)知

$$(A+\tilde{E})x=0 \text{ 且 } x\neq0,$$

从而 $\text{rank}(A+\tilde{E})<n$。因而，对任给的 $\varepsilon>0$，存在 $D_\varepsilon\in\mathbb{R}^{m\times n}$ 满足 $\|D_\varepsilon\|_F<\varepsilon$ 使得

$$b+\tilde{r}\in\mathscr{R}(A+\tilde{E}+D_\varepsilon).$$

因此，

$$\begin{aligned}d&\leqslant\|[\tilde{E}+D_\varepsilon,\tilde{r}]\|_F\\&\leqslant\|[\tilde{E},\tilde{r}]\|_F+\|D_\varepsilon\|_F\\&\leqslant\sigma_{n+1}+\varepsilon.\end{aligned}$$

注意到 ε 的任意性，结合(5.6)即知 $d=\sigma_{n+1}$，即此时亦有引理成立。证毕。

定理5.1 TLS 问题(5.1)有解的充分必要条件是至少存在一个对应于 $[A,b]$ 的最小奇异值 σ_{n+1} 的右奇异向量 v，其最后一个分量不为零。

证明 充分性 设 v 是对应于 σ_{n+1} 的右奇异向量，而且 v 的最后一个分量不为零。现取 u 是对应于 v 属于 σ_{n+1} 的左奇异向量，即

$$[A,b]v=\sigma_{n+1}u.$$

当然，我们亦可要求 $\|v\|_2=\|u\|_2=1$。

令

$$[E_0, r_0] = -\sigma_{n+1} u v^{\mathrm{T}}.$$

则有

$$\|[E_0, r_0]\|_F = \sigma_{n+1}, \tag{5.8}$$

$$[A+E_0, b+r_0]v = \sigma_{n+1} u - \sigma_{n+1} u = 0. \tag{5.9}$$

注意到 v 的最后一个分量不为零，从(5.9)即知

$$b + r_0 \in \mathscr{R}(A + E_0).$$

由引理 5.3 即知$[E_0, r_0]$是 TLS 问题(5.1)的一个解.

必要性 设 TLS 问题(5.1)有解$[E, r]$. 则由引理 5.3 知，

$$\|[E, r]\|_F = \sigma_{n+1}.$$

再由 $b + r \in \mathscr{R}(A+E)$知，必存在 $x \in \mathbb{R}^n$ 使得

$$(A+E)x = b + r.$$

现令

$$v = \begin{bmatrix} x \\ -1 \end{bmatrix}.$$

对于矩阵$[A, b]$,$[E, r]$和向量 v 应用引理 5.2,即知 v 就是$[A, b]$之属于 σ_{n+1} 的右奇异向量. 显然， v 的最后一个分量不为零.

注5.1 设 $v = (\zeta_1, \cdots, \zeta_n, \zeta_{n+1})^{\mathrm{T}}$ 是$[A, b]$之属于最小奇异值 σ_{n+1} 的右奇异向量，且 $\zeta_{n+1} \neq 0$. 则由定理 5.1 的证明知，

$$x = -\left(\frac{\zeta_1}{\zeta_{n+1}}, \frac{\zeta_2}{\zeta_{n+1}}, \cdots, \frac{\zeta_n}{\zeta_{n+1}}\right)^{\mathrm{T}} \tag{5.10}$$

是对应的 TLS 问题(5.1)的一个 TLS 解；反之，若 x 是 TLS问题(5.1)的一个 TLS 解，则 $v = \begin{bmatrix} x \\ -1 \end{bmatrix}$ 就是对应于 σ_{n+1} 的一个右奇异向量.

记

$$\mathscr{X}_{\mathrm{TLS}} = \{x \in \mathbb{R}^n:\ x \text{ 是问题(5.1)的 TLS 解}\}.$$

则有如下结果：

推论5.1 设 TLS 问题(5.1)有解。则有:

(1) $\mathscr{X}_{\mathrm{TLS}}$ 是非空凸集;

(2) 有且仅有唯一的 $x_{\mathrm{TLS}}\in\mathscr{X}_{\mathrm{TLS}}$,使得

$$\|x_{\mathrm{TLS}}\|_2=\min\{\|x\|_2:\ x\in\mathscr{X}_{\mathrm{TLS}}\},\qquad(5.11)$$

通常称 x_{TLS} 为问题(5.1)的最小范数 TLS 解;

(3) $\mathscr{X}_{\mathrm{TLS}}=\{x_{\mathrm{TLS}}\}$的充分必要条件是$[A,b]$的最小奇异值 σ_{n+1} 是单重的。

证明 (1) 由问题(5.1)有解,故$\mathscr{X}_{\mathrm{TLS}}$非空是显然的。现假定 $x,y\in\mathscr{X}_{\mathrm{TLS}}$,则由注 5.1 知$\begin{bmatrix}x\\-1\end{bmatrix}$和$\begin{bmatrix}y\\-1\end{bmatrix}$都是$[A,b]$之对应于 σ_{n+1} 的右奇异向量,从而对任意的 $t\in[0,1]$, $t\begin{bmatrix}x\\-1\end{bmatrix}+(1-t)\begin{bmatrix}y\\-1\end{bmatrix}$亦是对应于 σ_{n+1}的右奇异向量,因此$tx+(1-t)y\in\mathscr{X}_{\mathrm{TLS}}$,即$\mathscr{X}_{\mathrm{TLS}}$是凸集。于是(1)得证。

(2) 如果有两个向量 $x,y\in\mathscr{X}_{\mathrm{TLS}}$,使得

$$\|x\|_2=\|y\|_2=\min\{\|x\|:\ x\in\mathscr{X}_{\mathrm{TLS}}\},\quad x\neq y,$$

那么,一方面由$\mathscr{X}_{\mathrm{TLS}}$凸知,对任意的 $t\in[0,1]$有 $tx+(1-t)y\in\mathscr{X}_{\mathrm{TLS}}$,从而

$$\|x\|_2\leqslant\|tx+(1-t)y\|_2,\quad\forall\, t\in[0,1];$$

另一方面,

$$\|tx+(1-t)y\|_2\leqslant t\|x\|_2+(1-t)\|y\|_2=\|x\|_2,\quad\forall\, t\in[0,1];$$

因此,对任意的 $t\in[0,1]$,有

$$\|tx+(1-t)y\|_2=\|x\|_2,$$

这与 $x\neq y$ 矛盾,故只有唯一的 $x\in\mathscr{X}_{\mathrm{TLS}}$ 使得(5.11)成立。即(2)得证。

(3) 若 σ_{n+1} 是单重的,则有且仅有唯一的$\begin{bmatrix}x\\-1\end{bmatrix}\in\mathbb{R}^{n+1}$,

使得$\begin{bmatrix} x \\ -1 \end{bmatrix}$是对应于$\sigma_{n+1}$的右奇异向量，故$\mathscr{X}_{TLS}$只含有唯一的点（注意，这里用到了TLS问题(5.1)有解的假定和定理5.1)。反之，若σ_{n+1}非单重的，则至少存在属于σ_{n+1}的两个线性无关的右奇异向量u和v。而问题(5.1)又是有解的，因此不妨设$v=\begin{bmatrix} x \\ -1 \end{bmatrix}$。如果$u$的最后一个分量亦不为零，则$u$亦可取为$\begin{bmatrix} y \\ -1 \end{bmatrix}$的形式，这样$\mathscr{X}_{TLS}$至少含有两个不同的点$x$和$y$；如果$u$的最后一个分量为0，则$u=\begin{bmatrix} y \\ 0 \end{bmatrix}$，而$y\neq 0$，从而$u+v=\begin{bmatrix} x+y \\ -1 \end{bmatrix}$亦是属于$\sigma_{n+1}$的一个右奇异向量，于是$x+y$和$x$都是$\mathscr{X}_{TLS}$中的点，而且$x\neq x+y$。这样就证明了(3)。证毕。

推论5.2 设σ_{n+1}和$\hat{\sigma}_n$分别是$[A,b]$和A的最小奇异值。如果$\hat{\sigma}_n>\sigma_{n+1}$，则对应的问题(5.1)有且只有唯一的TLS解。

证明 设$v=\begin{bmatrix} y \\ \zeta_{n+1} \end{bmatrix}\begin{matrix} n \\ 1 \end{matrix}$是$[A,b]$之属于$\sigma_{n+1}$的右奇异向量。则必有$\zeta_{n+1}\neq 0$。若不然，假定$u$是对应的左奇异向量，即

$$[A,b]v=\sigma_{n+1}u.$$

当然，亦可要求$\|v\|_2=\|u\|_2=1$。这样由$\zeta_{n+1}=0$，即知

$$Ay=\sigma_{n+1}u, \qquad \|y\|_2=\|v\|_2=1,$$

从而有

$$\hat{\sigma}_n=\min_{\|x\|_2=1}\|Ax\|_2\leqslant\|Ay\|_2\leqslant\sigma_{n+1},$$

这与已知条件$\sigma_{n+1}<\hat{\sigma}_n$矛盾。因此，$v$的最后一个分量不为零，从而由定理5.1即知对应的问题(5.1)有解。

由于

$$[A,b]^{T}[A,b]=\begin{bmatrix} A^{T}A & A^{T}b \\ b^{T}A & b^{T}b \end{bmatrix},$$

利用第一章的推论6.5即知，条件 $\sigma_{n+1}<\hat{\sigma}_n$ 蕴含着 σ_{n+1} 是$[A,b]$的单重奇异值，从而由推论5.1的(3)知它有唯一的 TLS解. 证毕.

现在，我们再来考虑如何求问题(5.1)的最小范数 TLS 解 x_{TLS}.

假设$[A,b]$的奇异值分解为

$$[A,b]=U\begin{bmatrix}\Sigma\\0\end{bmatrix}V^{\mathrm{T}}, \tag{5.12}$$

其中 $U\in\mathbb{R}^{m\times m}$ 正交矩阵，$V=[v_1,v_2,\cdots,v_{n+1}]\in\mathbb{R}^{(n+1)\times(n+1)}$ 正交矩阵，$\Sigma=\mathrm{diag}(\sigma_1,\cdots,\sigma_{n+1})$，$\sigma_1\geqslant\cdots\geqslant\sigma_p>\sigma_{p+1}=\sigma_{p+2}=\cdots=\sigma_{n+1}$. 令

$$\mathscr{S}=\{v\in\mathbb{R}^{n+1}:\ \|v\|_2^2=1\}\cap\mathrm{span}\{v_{p+1},\cdots,v_{n+1}\}.$$

若 TLS 问题(5.1)有解，则由注5.1知

$$\mathscr{X}_{\mathrm{TLS}}=\left\{x\in\mathbb{R}^n:\ \frac{1}{\|x\|_2^2+1}\begin{bmatrix}x\\-1\end{bmatrix}\in\mathscr{S}\right\}.$$

从而，如果 v 是$\mathscr{S}$中最后一个分量最大者，则按(5.10)构造出来的 x 必是问题(5.1)的最小范数 TLS 解 x_{TLS}.

现取一个 $n+1-p$ 阶 Householder 矩阵H，使得

$$[v_{p+1},\cdots,v_{n+1}]H=\begin{bmatrix}X & z\\0 & \alpha\end{bmatrix}\begin{matrix}n\\1\end{matrix},\quad \alpha\geqslant0. \tag{5.13}$$
$$\qquad\qquad\qquad\qquad n-p\quad 1$$

则易证 $v=\begin{bmatrix}z\\\alpha\end{bmatrix}$就是$\mathscr{S}$中最后一个分量最大者. 如果 $\alpha=0$，则 TLS 问题(5.1)无解；如果 $\alpha\neq0$，则

$$x_{\mathrm{TLS}}=-\frac{1}{\alpha}z. \tag{5.14}$$

综上所述，可得如下算法.

算法5.1

(1) 输入 A,b。

(2) 计算$[A,b]$的奇异值分解(5.12)。

(3) 确定下标 p 使得 $\sigma_1\geqslant\cdots\geqslant\sigma_p>\sigma_{p+1}=\cdots=\sigma_{n+1}$。

(4) 确定 Householder 变换H使(5.13)成立；如果 $\alpha=0$，则 TLS 问题(5.1)无解，结束；否则进行下一步。

(5) 按照(5.14)计算 x_{TLS}。

(6) 输出 x_{TLS}，结束。

此一算法中的奇异值分解用第七章的算法5.2来计算，这样，它的运算量为 $2mn^2+12n^3$。

习　题

1. 证明：对任意的 $A\in\mathbb{R}^{m\times n}$ 有 $\mathscr{N}(A^{\mathrm{T}}A)=\mathscr{N}(A)$。

2. 利用正规化方法求最小二乘问题(1.1)的解，其中

$$A=\begin{bmatrix}1&2\\3&4\\5&6\end{bmatrix},\quad b=\begin{bmatrix}1\\1\\1\end{bmatrix}.$$

3. 证明恒等式

$$\|A(x+\alpha w)-b\|_2^2=\|Ax-b\|_2^2-2\alpha w^{\mathrm{T}}A^{\mathrm{T}}(Ax-b)+\alpha^2\|Aw\|_2^2,$$

其中 $A\in\mathbb{R}^{m\times n}$，$b\in\mathbb{R}^m$，$x,w\in\mathbb{R}^n$，$\alpha\in\mathbb{R}$。并利用这一恒等式给出定理1.1的另一种证明。

4. 给出用 Givens 变换求解满秩 LS 问题的详细算法。

5. 证明：如果 $A_k\to A$ 且 $A_k^{\dagger}\to A^{\dagger}$，则存在一个自然数 k_0，使当 $k\geqslant k_0$ 时，$\mathrm{rank}(A_k)$等于常数。

6. 证明：如果 $A\in\mathbb{R}^{m\times n}$ 的秩是 n，则对任意的 $E\in\mathbb{R}^{m\times n}$，只要 $\|A^{\dagger}\|_2\|E\|_2<1$，就有 $A+E$ 的秩亦等于 n。

7. 设

$$R=\begin{bmatrix}R_{11}&R_{12}\\0&R_{22}\end{bmatrix}\begin{matrix}k\\m-k\end{matrix}\,,$$
$$\qquad\begin{matrix}k\quad&n-k\end{matrix}$$

证明：$\sigma_{k+1}(R) \leqslant \|R_{22}\|_2$，其中 $\sigma_{k+1}(R)$表示R的第 $k+1$ 个奇异值。

8．证明：方程

$$\omega^4(2-\omega)^3 = (1-\omega) + (1-\omega)^6$$

在 $0<\omega<1$ 内有唯一的解 $\omega_0 = \frac{1}{2}(\sqrt{5}-1)$。

9．设 $A \in \mathbb{R}^{m \times n}$，$v \in \mathbb{R}^n$，$\|v\|_2 = 1$，满足$\|Av\|_2 = \varepsilon$。并假定 P 是一排列方阵，它使得 $P^{\mathrm{T}}v = w = (\omega_1, \cdots, \omega_n)^{\mathrm{T}}$ 满足$|\omega_n| = \|w\|_\infty = \|v\|_\infty$。证明：如果$AP = QR$ 是 AP 的 QR 分解，则R的第 n 个对角元素 r_{nn} 满足

$$|r_{nn}| < n^{\frac{1}{2}}\varepsilon.$$

利用这一结果设计一种判定 A 的数值秩的数值方法。

10．讨论如下的约束最小二乘问题：

$$\min_x \|Ax - b\|_2,$$

$$\text{s.t.} \quad Bx = d,$$

其中 $A \in \mathbb{R}^{m \times n}$，$B \in \mathbb{R}^{p \times n}$，$b \in \mathbb{R}^m$，$d \in \mathbb{R}^p$，$x \in \mathbb{R}^n$。并设计一种求解的数值方法。

第七章　求解特征值问题的QR方法

§1　特征值和不变子空间的条件数

这一节，我们来考虑特征值问题对扰动的敏感程度的定量刻画。

矩阵元素的微小扰动将会对其特征值和特征向量产生什么样的影响是一个十分复杂的问题。它不仅与扰动量的大小有关，而且亦与扰动的方式有关。例如，矩阵

$$A=\begin{bmatrix} 0 & 1 & & 0 \\ & 0 & \ddots & \\ & & \ddots & \\ & 0 & & 1 \\ & & & 0 \end{bmatrix}\in \mathbb{R}^{n\times n}$$

有 $\lambda=0$ 作为其 n 重特征值，对 A 作微小扰动变为

$$\tilde{A}=A+\varepsilon e_n e_1^{\mathrm{T}},\quad 0<\varepsilon\ll 1,$$

则 $\tilde{A}$ 有 n 个互不相同的特征值 $\lambda_1,\cdots,\lambda_n$，且满足

$$|\lambda_i|=\varepsilon^{1/n},\quad i=1,2,\cdots,n.$$

这表明，对矩阵元素的 ε 量级的变化，可引起特征值 $\varepsilon^{1/n}$ 量级的变化。

其次，A 的特征子空间是一维的，而 $\tilde{A}$ 有 n 个线性无关的特征向量。因此，其特征向量的结构发生了本质上的变化。

如果对 A 作如下扰动

$$\tilde{A}=A+\delta e_1 e_n^{\mathrm{T}}.$$

则不论 δ 多大，$\tilde{A}$ 仍以 0 作为其 n 重特征值，而且其对应的特征子空间仍是一维的。

因此，要对特征值问题进行细致深入的扰动分析，就需花费

大量的时间，占用大量的篇幅。这里，由于篇幅所限，只能对一些最基本的结果做一简要的介绍。

1.1 特征值的条件数

第一章曾介绍的 Weilandt-Hoffman 定理表明，Hermite矩阵和正规矩阵的特征值是良态的，即特征值对矩阵元素的微小扰动不敏感。因此，我们这里着重需要讨论非正规矩阵的特征值对扰动的敏感程度的数据标准。

设 A 是一个 n 阶方阵，并假定 A 的 Jordan分解为

$$Q^{-1}AQ = J, \tag{1.1}$$

其中 Q 是 n 阶非奇异方阵，J 是 A 的 Jordan 标准形。利用广义 Bauer-Fike 定理易证：对任意的 $E\in\mathbb{C}^{n\times n}$，只要

$$\|QEQ^{-1}\|_2\leqslant 2^{1-p},$$

就有对任意的 $\mu\in\lambda(A+E)$，必存在 $\lambda\in\lambda(A)$使得

$$|\mu-\lambda|\leqslant 2^{1-1/p}(\|Q\|_2\|Q^{-1}\|_2)^{1/p}\|E\|_2^{1/p}, \tag{1.2}$$

其中 p 是 A 的 Jordan 标准形中最大 Jordan 块的阶数。

(1.2) 表明，量$\|Q\|_2\|Q^{-1}\|_2$ 在一定程度上反映了 A 的特征值对 A 的元素的微小扰动的敏感程度。因此，通常称数

$$v(A)=\inf_{Q\in\mathscr{D}_A}\|Q\|_2\|Q^{-1}\|_2 \tag{1.3}$$

为 A 的**谱条件数**，其中

$$\mathscr{D}_A=\{Q\in\mathbb{C}^{n\times n}:\ Q^{-1}AQ=J\}.$$

显然有 $v(A)\geqslant 1$，而且当 A 是正规矩阵时，有 $v(A)=1$。

虽然(1.3)所定义的谱条件数 $v(A)$的大小在一定程度上反映了其特征值对扰动的敏感程度，但它是对 A 的全部特征值整体而言的，并不反映每个具体特征值的敏感程度。对于一个非正规方阵，它的特征值，有的可能对扰动十分敏感，而有的则不敏感。例如，矩阵

$$A=\begin{bmatrix} a & 0 & & 0 & \\ & b & 1 & & \\ & & \ddots & \ddots & \\ 0 & & & \ddots & 1 \\ & & & & b \end{bmatrix}\in \mathbb{C}^{n\times n},\quad a\neq b,$$

有两个互不相同的特征值 a 和 b。对 A 作如下的微小扰动

$$\tilde{A}=A+\varepsilon e_n e_2^{\mathrm{T}},\quad 0<\varepsilon\ll 1,$$

则特征值 a 不变，而特征值 b 变为 $n-1$ 个互不相同的特征值

$$\lambda_j=b+\varepsilon^{\frac{1}{n-1}}e^{\frac{2\pi j}{n-1}i},\quad j=0,1,\cdots,n-2.$$

因此，我们需对每个特征值分别对待，给出其条件数的定义。

先来考虑单特征值的情形。设 λ 是 A 的单特征值，x 和 y 分别是 A 之属于 λ 的右、左特征向量，即

$$Ax=\lambda x,\quad y^{\mathrm{T}}A=\lambda y^{\mathrm{T}},\quad x\neq 0,\ y\neq 0.$$

现不妨假定 $\|x\|_2=\|y\|_2=1$。则容易证明数

$$s(\lambda)=|y^{\mathrm{T}}x| \tag{1.4}$$

是由 λ 唯一确定的，而且 $s(\lambda)>0$。

此外，利用函数论的一些基本结果，可证：对任意的 $E\in\mathbb{C}^{n\times n}$，存在 $\varepsilon_0>0$ 以及在 $|\varepsilon|\leqslant\varepsilon_0$ 上定义的解析函数 $\lambda(\varepsilon)$，满足：

(1) 在 $|\varepsilon|\leqslant\varepsilon_0$ 上，$\lambda(\varepsilon)$ 是矩阵 $A+\varepsilon E$ 的单特征值；

(2) $\lambda(0)=\lambda$，$\lambda'(0)=y^{\mathrm{T}}Ex/y^{\mathrm{T}}x$。 (1.5)

上述结论的详细证明参见文献[10]。

从(1.5)可知，当 $\|E\|_2\leqslant 1$ 时，有

$$|\lambda'(0)|=\left|\frac{y^{\mathrm{T}}Ex}{y^{\mathrm{T}}x}\right|\leqslant\frac{1}{|y^{\mathrm{T}}x|}=\frac{1}{s(\lambda)}, \tag{1.6}$$

且当 $E=\bar{y}x^*$ 时，达到上界。因此，有

$$|\lambda(\varepsilon)-\lambda|=|\lambda'(0)\varepsilon+O(\varepsilon^2)|\leqslant\frac{\varepsilon}{s(\lambda)}+O(\varepsilon^2). \tag{1.7}$$

粗略地讲，(1.7)表明，A的元素有ε量级的扰动，则其特征值λ的变化大约为$\varepsilon/s(\lambda)$。因此，我们称数$1/s(\lambda)$为A的单特征值λ的**条件数**．当$s(\lambda)$很小时，就说λ是**病态的**；否则，就说λ是**良态的**。

另外，Wilkinson于1972年从几何上阐明了$s(\lambda)$很小的实质是A与一个以λ为其重特征值的矩阵很靠近．他证明了当$s(\lambda)<1$时，存在E满足

$$\|E\|_2\leqslant s(\lambda)[1-s(\lambda)^2]^{-1/2},$$

使得$A+E$以λ为其重特征值。

当λ是A的重特征值时，其敏感性问题变得非常复杂，至今仍没有得到彻底解决。只是当λ非亏损时，孙继广于1992年给出了其条件数的定义，但其讨论相当复杂，因此这里不再赘述，有兴趣的读者可参阅文献[65]。

1.2 不变子空间的条件数

下面我们再来考虑特征向量的敏感性问题。先假定$A\in\mathbb{C}^{n\times n}$有$n$个互异的特征值$\lambda_1,\cdots,\lambda_n$，对应的左右单位特征向量分别为$y_1,\cdots,y_n$和$x_1,x_2,\cdots,x_n$。再设$E\in\mathbb{C}^{n\times n}$满足$\|E\|_2=1$。则利用函数论的某些结果可证，存在$\varepsilon_0>0$和在$|\varepsilon|<\varepsilon_0$上解析的数值函数$\lambda_i(\varepsilon)$和向量值函数$x_i(\varepsilon)$，$i=1,2,\cdots,n$，使得

$$(A+\varepsilon E)x_i(\varepsilon)=\lambda_i(\varepsilon)x_i(\varepsilon), \tag{1.8}$$

并满足

$$\lambda_i(0)=\lambda_i,\quad x_i(0)=x_i.$$

假定$x_i(\varepsilon)$的泰勒展式为

$$x_i(\varepsilon)=x_i+\varepsilon z_i+O(\varepsilon^2), \tag{1.9}$$

由于$x_1,\cdots,x_n$作成$\mathbb{C}^n$的一组基，故有

$$z_i=\sum_{j=1}^{n}a_{ij}x_j,\quad a_{ij}\in\mathbb{C}. \tag{1.10}$$

将(1.10)代入(1.9)即有

$$x_i(\varepsilon)=(1+\varepsilon\alpha_{ii})x_i+\varepsilon\sum_{j\neq i}\alpha_{ij}x_j+O(\varepsilon^2). \qquad (1.11)$$

由此可见，我们可以通过给 $x_i(\varepsilon)$适当地乘以一个解析的数值函数，使得对应的泰勒展式中 z_i 的表达式(1.10)中之 $\alpha_{ii}=0$。现在，假定 $x_i(\varepsilon)$已选好使得 $\alpha_{ii}=0$，并将 $\lambda_i(\varepsilon)$的泰勒展式

$$\lambda_i(\varepsilon)=\lambda_i+\lambda_i'(0)\varepsilon+O(\varepsilon^2)$$

和(1.11)代入(1.8)，再比较 ε 项之系数，可得

$$\sum_{j\neq i}\alpha_{ij}(\lambda_j-\lambda_i)x_j=(\lambda_i'(0)I-E)x_i. \qquad (1.12)$$

在(1.12)两边左乘 y_j^{T}，并注意到 $y_j^{\mathrm{T}}x_i=0$，$i\neq j$，即有

$$\alpha_{ij}(\lambda_j-\lambda_i)y_j^{\mathrm{T}}x_j=-y_j^{\mathrm{T}}Ex_i,\quad j\neq i,$$

即

$$\alpha_{ij}=y_j^{\mathrm{T}}Ex_i/[(\lambda_i-\lambda_j)y_j^{\mathrm{T}}x_j],\quad j\neq i, \qquad (1.13)$$

无妨假定 $s(\lambda_j)=|y_j^{\mathrm{T}}x_j|=y_j^{\mathrm{T}}x_j$。则将(1.13)代入(1.11)即得

$$x_i(\varepsilon)=x_i+\varepsilon\sum_{j\neq i}\frac{y_j^{\mathrm{T}}Ex_i}{(\lambda_i-\lambda_j)s(\lambda_j)}\cdot x_j+O(\varepsilon^2). \qquad (1.14)$$

这表明，特征向量的敏感程度，不仅与其他特征值的条件数的大小有关，而且亦与这一特征向量所对应的特征值与其他特征值的分离程度有关。当特征值分离不明显时，特征向量对扰动就会变得十分敏感。因此，相对集中的特征值，对应的特征向量是十分病态的。但是，如果将敏感的特征向量放在一起作成 A 的不变子空间，则可以是不敏感的。为了说明这一事实，我们先简要地回顾一下不变子空间的定义和基本性质。

设 $A\in\mathbb{C}^{n\times n}$，$\mathscr{X}\subset\mathbb{C}^n$ 是一个 l 维子空间。如果 $A\mathscr{X}\subset\mathscr{X}$（即对任意的 $x\in\mathscr{X}$ 有 $Ax\in\mathscr{X}$），则称$\mathscr{X}$是 A 的一个**不变子空间**。

设$\mathscr{X}=\mathscr{R}(X_1)$，其中 $X_1\in\mathbb{C}^{n\times l}$，$X_1^*X_1=I_l$。则由不变子空间的定义易证，$\mathscr{X}$是 A 的不变子空间的充分必要条件是存在

$A_{11}\in\mathbb{C}^{l\times l}$ 使得

$$AX_1 = X_1A_{11}. \qquad (1.15)$$

现假定$\mathscr{X}=\mathscr{R}(X_1)(X_1\in\mathbb{C}^{n\times l}, X_1^*X_1=I_l)$ 是A的一个不变子空间，并假定 $A_{11}\in\mathbb{C}^{l\times l}$ 满足(1.15)。再设 $\lambda(A_{11})=\{\lambda_1,\cdots,\lambda_l\}$。则易知 $\lambda_1,\cdots,\lambda_l$ 亦是A的特征值，且由$\mathscr{X}$唯一确定与X_1的选取无关。因此，通常称 $\lambda_1,\cdots,\lambda_l$ 为A在不变子空间$\mathscr{X}$上的限制 $A|_{\mathscr{X}}$ 的特征值。

设 $\lambda(A)=\{\lambda_1,\cdots,\lambda_n\}$，且$\{\lambda_1,\cdots,\lambda_l\}\cap\{\lambda_{l+1},\cdots,\lambda_n\}=\varnothing$。则有且仅有唯一的$A$的不变子空间$\mathscr{X}$使得 $A|_{\mathscr{X}}$ 的特征值就是 $\lambda_1,\cdots,\lambda_l$。即由 $\lambda_1,\cdots,\lambda_l$ 可确定A的唯一的 l 维不变子空间。

事实上，由 Schur 分解定理知，存在酉矩阵Q使得

$$Q^*AQ=T, \qquad (1.16)$$

其中T是对角元素依次为 $\lambda_1,\cdots,\lambda_n$ 的上三角矩阵。现将Q和T分块如下

$$Q=\underset{\quad l \qquad n-l}{[Q_1,\ Q_2]},\qquad T=\begin{bmatrix}T_{11} & T_{12}\\ 0 & T_{22}\end{bmatrix}\begin{matrix}l\\ n-l\end{matrix},$$
$$\qquad\qquad\qquad\qquad\quad l \qquad n-l$$

则由(1.16)，可得

$$AQ_1=Q_1T_{11}, \qquad (1.17)$$

$$Q_2^*A=T_{22}Q_2^*. \qquad (1.18)$$

从(1.17)即知，$\mathscr{X}=\mathscr{R}(Q_1)$是$A$的一个 l 维不变子空间，而且 $A|_{\mathscr{X}}$ 的特征值正好就是 $\lambda_1,\cdots,\lambda_l$。

若$\mathscr{X}_2=\mathscr{R}(X_1)(X_1\in\mathbb{C}^{n\times l}, X_1^*X_1=I_l)$也是$A$的一个不变子空间，并且 $A|_{\mathscr{X}_2}$ 的特征值也是 $\lambda_1,\lambda_2,\cdots,\lambda_l$，则存在 $A_{11}\in\mathbb{C}^{l\times l}$ 使得

$$AX_1=X_1A_{11}, \qquad (1.19)$$

而且 A_{11} 的特征值是 $\lambda_1,\cdots,\lambda_l$。由(1.18)和(1.19)可得

$$T_{22}Q_2^*X_1=Q_2^*AX_1=Q_2^*X_1A_{11}$$

即

$$T_{22}(Q_2^* X_1) - (Q_2^* X_1) A_{11} = 0。 \tag{1.20}$$

而条件 $\lambda(T_{22}) \cap \lambda(A_{11}) = \{\lambda_{l+1}, \cdots, \lambda_n\} \cap \{\lambda_1, \cdots, \lambda_l\} = \varnothing$，蕴含着矩阵方程组

$$T_{22} Y - Y A_{11} = 0$$

只有唯一的零解 $Y=0$。故(1.20)蕴含着

$$Q_2^* X_1 = 0。$$

从而有

$$\mathscr{X}_2 = \mathscr{R}(X_1) = \mathscr{R}(Q_2)^{\perp} = \mathscr{R}(Q_1) = \mathscr{X}。$$

如果 A 的特征值满足

$$|\lambda_1| \geqslant \cdots \geqslant |\lambda_l| > |\lambda_{l+1}| \geqslant \cdots \geqslant |\lambda_n|,$$

则称由 $\lambda_1, \cdots, \lambda_l$ 所确定的唯一的不变子空间为 A 的**优势不变子空间**，通常记作 $\mathscr{D}_l(A)$。

现在我们再来考虑不变子空间对扰动的敏感性问题。从前面对特征向量敏感性的粗略分析可知，对于 A 的一个不变子空间 $\mathscr{X}$，其敏感程度应与 $A|_{\mathscr{X}}$ 的特征值与 A 的其余特征值的分离程度有关，而这种分离程度的较为合适的度量就是如下定义的分离度：

设 $B \in \mathbb{C}^{l \times l}$，$C \in \mathbb{C}^{m \times m}$。则称数

$$\operatorname{sep}(B, C) = \inf_{\substack{P \in \mathbb{C}^{m \times l} \\ \|P\|_2 = 1}} \|PB - CP\|_2 \tag{1.21}$$

为 **B 与 C 的分离度**。

可以证明，当 $\lambda(B) \cap \lambda(C) = \varnothing$ 时，$\operatorname{sep}(B, C) > 0$；而且对任意的 B 和 C 有

$$\operatorname{sep}(B, C) \leqslant \min\{|\lambda - \mu|: \lambda \in \lambda(B), \mu \in \lambda(C)\}。$$

由此可知，分离度确实在一定程度上反映了 B 与 C 的谱集之间的远近程度。

利用分离度的概念，我们可以给出不变子空间扰动上界的如下估计：

设 $A,E\in\mathbb{C}^{n\times n}$，$X=[X_1,X_2]$为酉矩阵，$X_1\in\mathbb{C}^{n\times l}(1\leqslant l\leqslant n-1)$。并假定$\mathscr{R}(X_1)$为$A$的一个不变子空间。再设 X^*AX和X^*EX 与X相一致地分块为

$$X^*AX=\begin{bmatrix}A_{11} & A_{12}\\ 0 & A_{22}\end{bmatrix},\qquad X^*EX=\begin{bmatrix}E_{11} & E_{12}\\ E_{21} & E_{22}\end{bmatrix}.$$

如果

$$\delta=\operatorname{sep}(A_{11},A_{22})-(\|E_{11}\|_2+\|E_{22}\|_2)>0,$$

并且

$$\delta^{-2}\|E_{21}\|_2(\|A_{12}\|_2+\|E_{12}\|_2)<\frac{1}{4},$$

则存在 $\tilde{A}=A+E$ 的不变子空间 $\mathscr{R}(\tilde{X}_1)$，$\tilde{X}_1\in\mathbb{C}_l^{n\times l}$，使得

$$\operatorname{dist}(\mathscr{R}(X_1),\mathscr{R}(\tilde{X}_1))\leqslant\frac{2\|E_{21}\|_2}{\delta}.\qquad(1.22)$$

上述结果的证明较繁，这里不再给出，有兴趣的读者可参看文献[10]的第三章第八节。

(1.22)表明：不变子空间敏感程度与δ有关。而对于微小扰动，$\delta\approx\operatorname{sep}(A_{11},A_{22})$。因此，通常用 $1/\operatorname{sep}(A_{11},A_{22})$作为不变子空间$\mathscr{R}(X_1)$的**条件数**。当 $\operatorname{sep}(A_{11},A_{22})$很小时，就说$\mathscr{R}(X_1)$ 是**病态的**；否则就说$\mathscr{R}(X_1)$是**良态的**。

作为本节的结束，我们再来看一个简单的例子。

设

$$A=\begin{bmatrix}A_{11} & 0\\ 0 & A_{22}\end{bmatrix},\quad 其中A_{11}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix},\quad A_{22}=\begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix}.$$

则容易算出，$\operatorname{sep}(A_{11},A_{22})=1$，且$A$之由特征值 1,1确定的唯一的不变子空间是$\mathscr{R}(X_1)$，其中

$$X_1=\begin{bmatrix}1 & 0\\ 0 & 1\\ 0 & 0\\ 0 & 0\end{bmatrix}.$$

如果对 A 作微小扰动变为

$$\tilde{A} = A + E = \begin{bmatrix} A_{11} & 0 \\ E_{21} & A_{22} \end{bmatrix},$$

其中

$$E_{21} = \begin{bmatrix} 0 & 0 \\ \varepsilon & 0 \end{bmatrix}, \quad 0 < \varepsilon \ll 1.$$

则通过简单的计算可知，$\tilde{A}$ 对应于特征值 1,1 的不变子空间是 $\mathscr{R}(\tilde{X}_1)$，其中

$$\tilde{X}_1 = \begin{bmatrix} \frac{1}{\sqrt{1+\varepsilon^2}} & 0 \\ 0 & 1 \\ 0 & 0 \\ \frac{-\varepsilon}{\sqrt{1+\varepsilon^2}} & 0 \end{bmatrix}.$$

因此，容易算出

$$\begin{aligned} \operatorname{dist}(\mathscr{R}(X_1), \mathscr{R}(\tilde{X}_1)) &= \|X_1 X_1^T - \tilde{X}_1 \tilde{X}_1^T\|_2 \\ &= \frac{\varepsilon}{\sqrt{1+\varepsilon^2}} \approx \frac{\|E_{21}\|_2}{\operatorname{sep}(A_{11}, A_{22})}. \end{aligned}$$

§2 双重步位移的QR算法

2.1 QR 算法的基本思想

QR 算法是电子计算机问世以来计算数学的重大进展之一，是目前计算中小型稠密矩阵全部特征值和特征向量的最有效的方法。它源于 Rutishauser 的 LR 算法，是由 Francis 和 Kublanovskaya 在60年代初期独立地提出的。其基本迭代格式如下：

$$\begin{cases} A_{m-1} = Q_m R_m, \\ A_m = R_m Q_m, \end{cases} \tag{2.1}$$

其中 $A_0=A\in\mathbb{C}^{n\times n}$，$Q_m$ 为酉矩阵，R_m 为上三角矩阵，$m=1, 2, \cdots$。

每个初次见到这一迭代的人，总会感到有点纳闷儿：这和特征值有什么关系？怎么会想到用这样的方法来求特征值呢？这一小节，我们就试图回答这些问题。为此，我们将从秩 1 矩阵（即秩为 1 的矩阵）的特征向量谈起。

设 $A=uv^{\mathrm{T}}$，其中 u，v 是 $\mathbb{C}^n$ 中两个非零向量。对任意的 $x\in\mathbb{C}^n$，$x\neq 0$，令

$$y=Ax=(v^{\mathrm{T}}x)u. \tag{2.2}$$

如果 $y=0$，则 x 就是属于 A 的零特征值的特征向量；如果 $y\neq 0$，则

$$\begin{aligned} Ay &= uv^{\mathrm{T}}y = uv^{\mathrm{T}}(v^{\mathrm{T}}x)u \\ &= (v^{\mathrm{T}}u)(v^{\mathrm{T}}x)u = (v^{\mathrm{T}}u)y, \end{aligned} \tag{2.3}$$

即 y 是对应于 A 的特征值 $v^{\mathrm{T}}u$ 的特征向量。

由此可见，如果我们事先只知道 A 是秩 1 矩阵，而并不知道其具体表达形式，则可从任一非零向量 x 出发，至多通过两次矩阵和向量的乘积，就可求得矩阵 A 的一个特征值和对应的特征向量。

对于一般的方阵 $A\in\mathbb{C}^{n\times n}$，假定 A 是非亏损的，即 A 有如下分解

$$A=X\Lambda Y^{\mathrm{T}}, \tag{2.4}$$

其中 $\Lambda=\mathrm{diag}(\lambda_1,\cdots,\lambda_n)$，$XY^{\mathrm{T}}=I$。用 x_i 和 y_i 分别表示 X 和 Y 的第 i 列，则（2.4）可改写为

$$A=\sum_{i=1}^{n}\lambda_i x_i y_i^{\mathrm{T}}. \tag{2.5}$$

再假定

$$|\lambda_1|>|\lambda_2|\geqslant\cdots\geqslant|\lambda_n|, \tag{2.6}$$

并记 $\tilde{A}=x_1 y_1^{\mathrm{T}}$，则有

$$\left\| \frac{A_k}{\lambda_1^k} - \tilde{A} \right\|_2 = \left\| \sum_{i=2}^{n} \left(\frac{\lambda_i}{\lambda_1} \right)^k x_i y_i^{\mathrm{T}} \right\|_2 \leqslant \eta \left| \frac{\lambda_2}{\lambda_1} \right|^k \longrightarrow 0, \quad k \to \infty,$$

其中 $\eta = (n-1)\max\limits_i \| x_i y_i^{\mathrm{T}} \|_2$。这表明，对于充分大的 k，A^k/λ_1^k 与秩 1 矩阵 $\tilde{A}$ 非常靠近。因此，由秩 1 矩阵的性质知，对任意的非零向量 x，当 k 充分大时，如果 $\frac{A^k}{\lambda_1^k}x \neq 0$，则向量

$$u_k = \frac{A^k}{\lambda_1^k} x \tag{2.7}$$

就是 A 的一个很好的近似特征向量。

这样，我们自然想到利用（2.7）来求 A 的近似特征向量。然而，实际计算时，这是行不通的。其原因有二：一是我们事先并不知道 A 的特征值 λ_1；二是对充分大的 k 计算 A^k 的工作量是大得惊人的。

仔细观察（2.7），不难发现（2.7）中的因子 λ_1^{-k} 仅对向量 $A^k x$ 的大小起作用，并不影响它的方向。而我们所感兴趣的只是 $A^k x$ 的方向，并非它的大小。因此，我们不必非用 λ_1^{-k} 来作为因子，而可用任意其他方便的常数作为因子（为了防止溢出，因子是必要的）。其次，计算 $A^k x$，也并不需事先将 A^k 算好之后再来计算，只需迭代地进行即可。基于这样的考虑，我们就可以设计如下的迭代格式：

$$\begin{aligned} & y_k = A u_{k-1}, \\ & \mu_k = \zeta_j^{(k)}, \qquad \zeta_j^{(k)} \text{ 是 } y_k \text{ 的模最大分量}, \\ & u_k = y_k/\mu_k, \qquad k = 1, 2, \cdots, \end{aligned} \tag{2.8}$$

其中 $u_0 \in \mathbb{C}^n$ 是任意给定的初始向量，$\|u_0\|_\infty = 1$。

易证在条件（2.6）成立的前提下，只要 u_0 选取得不要太糟糕，(2.8)产生的向量序列 $\{u_k\}_1^\infty$ 就收敛到 A 之属于 λ_1 的一个特征向量 v_1，数值序列 $\{\mu_k\}_{k=1}^\infty$ 就收敛到 λ_1，收敛速度依赖于 $\left|\frac{\lambda_2}{\lambda_1}\right|$ 的大小。按照这一迭代格式设计的算法称作**乘幂法**，目前

它仍是求大型稀疏矩阵少数几个模最大特征值常用的方法之一.

现在我们再来考察一下乘幂法的几何意义. 大家知道, 特征向量 v_1 只不过是特征子空间 $\mathrm{span}\{v_1\}$ 的一个代表, 而这个特征子空间正是我们感兴趣的对象. 类似地, 迭代所得到的每个 u_k 都可看作子空间 $\mathrm{span}\{u_k\}=\mathrm{span}\{A^k u_0\}$ 的一个代表. 这样一来, 乘幂法就可看作对子空间的迭代过程: 首先选取一个一维子空间 $S=\mathrm{span}\{u_0\}$, 然后逐步形成迭代序列

$$S, AS, \cdots, A^k S, \cdots, \tag{2.9}$$

这里 $A^k S$ 表示 S 在 A^k 作用之下的像空间. 子空间 $A^k S$ 将随着 k 的增大逐步逼近 A 对应于 λ_1 的特征子空间 $T=\mathrm{span}\{v_1\}$, 即

$$\mathrm{dist}(A^k S, T) \longrightarrow 0, \quad k \to \infty.$$

将乘幂法的这一基本思想推广到一般情形, 我们可以选取一个 l 维子空间 S, 并形成序列 (2.9). 可以想象, 这样得到的子空间序列应该收敛到 A 的一个 l 维不变子空间. 由于实际上, 我们不可能在整个子空间上进行迭代, 而是必须选取 S 的一组基, 在基上作同时迭代. 具体来讲, 实际上按如下格式进行迭代:

$$\begin{cases} Z_k = AQ_{k-1}, \\ Q_k R_k = Z_k \end{cases} \quad (k=1,2,\cdots), \tag{2.10}$$

其中 Q_0 是 S 的标准正交基作成的 $n\times l$ 矩阵, Q_k 是满足 $Q_k^* Q_k = I_l$ 的 $n\times l$ 矩阵, R_k 是 $l\times l$ 上三角矩阵. 易见 Q_k 的列就构成子空间 $A^k S$ 的一组标准正交基. 通常称迭代法 (2.10) 为正交迭代法, 有时亦称子空间迭代法或同时迭代法.

现假定 A 的特征值满足

$$|\lambda_1| \geqslant \cdots \geqslant |\lambda_l| > |\lambda_{l+1}| \geqslant \cdots \geqslant |\lambda_n|, \tag{2.11}$$

并假定 Q_0 满足

$$\mathscr{R}(Q_0) \cap \mathscr{D}_l(A^*)^\perp = \{0\}. \tag{2.12}$$

则可证存在常数 C, 使得

$$\text{dist}(\mathscr{R}(Q_k),\ \mathscr{D}_l(A)) \leqslant C\left|\frac{\lambda_{l+1}}{\lambda_l}\right|^k, \qquad (2.13)$$

其中$\mathscr{D}_l(A)$和$\mathscr{D}_l(A^*)$分别表示A和A^*的l维优势不变子空间。这一结果的证明十分繁锁，这里不再给出，有兴趣的读者可参看文献[37]。

因此，由（2.10）迭代产生的子空间序列$\{\mathscr{R}(Q_k)\}_{k=1}^{\infty}$将以速度$|\lambda_{l+1}/\lambda_l|^k$收敛于$A$的优势不变子空间$\mathscr{D}_l(A)$。

现在迭代（2.10）中取$l=n$，并假定A的特征值满足

$$|\lambda_1|>|\lambda_2|>\cdots>|\lambda_n|.$$

记

$$Q_k=[q_1^{(k)},\cdots,q_n^{(k)}], \qquad k=0,2,\cdots,$$

再假定

$$\text{span}\{q_1^{(0)},\cdots,q_i^{(0)}\}\cap\mathscr{D}_i(A^*)^{\perp}=\{0\}, \quad i=1,2,\cdots,n-1,$$

则有

$$\text{dist}(\mathscr{D}_i(A),\text{span}\{q_1^{(k)},\cdots,q_i^{(k)}\})\to 0, \qquad k\to\infty$$

对$i=1,2,\cdots,n-1$成立，其中的$\mathscr{D}_i(A)$和$\mathscr{D}_i(A^*)$分别表示A和A^*的i维优势不变子空间。

由此可见，如果定义

$$T_k=Q_k^*AQ_k, \qquad k=1,2,\cdots,$$

则T_k“逼近”于一个上三角矩阵。在这个意义下讲，$l=n$时的正交迭代实质上是逐次“逼近”A的Schur分解的一种迭代法。

从（2.10）可得

$$T_{k-1}=Q_{k-1}^*AQ_{k-1}=(Q_{k-1}^*Q_k)R_k,$$
$$T_k=Q_k^*AQ_k=Q_k^*AQ_{k-1}(Q_{k-1}^*Q_k)=R_k(Q_{k-1}^*Q_k).$$

因此，如果我们将（2.10）直接变为由T_{k-1}到T_k的迭代，就可得到如下的迭代：

$$T_0=A,$$
$$T_{k-1}=\tilde{Q}_kR_k \quad (\text{QR 分解}),$$

$$T_k = R_k \tilde{Q}_k, \qquad k = 1, 2, \cdots.$$

这就得到了著名的 QR 迭代。至此，大家已经明白，QR 迭代实质就是 $l = n$ 时正交迭代法的一种巧妙的实现，而正交迭代法是乘幂法的一种自然推广。

2.2 实 Schur 标准形

由于实际应用中所遇到的大量特征值问题都是关于实矩阵的，因此我们自然希望设计只涉及实运算的 QR 迭代，即给定 $A \in \mathbb{R}^{n \times n}$，令 $A_1 = A$，构造迭代：

$$\begin{cases} A_k = Q_k R_k, \\ A_{k+1} = R_k Q_k, \end{cases} \qquad k = 1, 2, \cdots, \tag{2.14}$$

其中 Q_k 是正交矩阵，R_k 是上三角矩阵。

然而，此时由于复共轭特征值的存在，我们自然不能期望（2.14）产生的 A_k 仍然逼近于一个上三角矩阵(即 A 的 Schur标准形)。那么，A_k 将趋向于什么呢？这就涉及到一个实矩阵在正交相似变换下的标准形问题。完全类似于 Schur 分解定理的证明可证：

定理2.1（实 Schur 分解） 设 $A \in \mathbb{R}^{n \times n}$。则存在正交矩阵 $Q \in \mathbb{R}^{n \times n}$，使得

$$Q^{\mathrm{T}} A Q = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1m} \\ 0 & R_{22} & \cdots & R_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & R_{mm} \end{bmatrix}, \tag{2.15}$$

其中 R_{ii} 或者是一个单元素，或者是一个具有一对复共轭特征值的 2 阶方阵。

这一定理的证明作为练习请读者自己给出。

（2.15）通常被称作实矩阵 A 的**实 Schur 分解**，而右边的拟上三角阵被称作 A 的**实 Schur 标准形**。显然，只要求得一个实矩阵的实 Schur 标准形，我们就可很容易求得它的全部特征值。

从前面关于正交迭代法收敛性的粗略分析，不难想象迭代（2.14）将逼近于A的实Schur分解。事实上也确实如此，在一定条件下，可证（2.14）产生的A_k将“逼近”于A的实Schur标准形。

其次（2.14）作为一种实用的迭代法是没有竞争力的，其原因有二：一是每次迭代的运算量太大（大约是$O(n^3)$）；二是收敛速度太慢（依赖于特征值的分离程度）。因此，要使其成为一种高效的方法，就必须尽可能地减少其每次迭代的运算量，提高其收敛速度。这就是本节下面所要讨论的主要内容。

在下面的讨论中，如无特别说明，我们总假定给定的n阶方阵A是实的。

2.3 上 Hessenberg 化

从（2.14）知，对一般的方阵完成一次QR迭代所需的运算量是$O(n^3)$。但是，如果在进行QR迭代之前，先对A进行适当的相似变换，使其具有较多的零元素，即适当选取非奇异矩阵Q_0，使

$$A_0 = Q_0^{-1} A Q_0 \tag{2.16}$$

具有较多的零元素，然后再对A_0进行QR迭代，则可望使每次迭代的运算量大为减少。

当然，我们只能利用矩阵计算的三种基本工具来实现Q_0和A_0的选取。首先来看利用Householder变换，可以得到什么样的A_0。

对于给定的$A=[a_{ij}]\in\mathbb{R}^{n\times n}$，第一步，我们自然应该选取Householder变换H_1，使H_1A的第一列有尽可能多的零元素（至多只能有$n-1$个零元素）。然而，为了保证对A进行相似变换，在对A进行了行变换之后，必须亦对A进行同样的列变换，即应将H_1亦右乘于H_1A上变为

$$H_1AH_1.$$

这样，为了保证已在H_1A的第一列所出现的零元素不致于在右

乘 H_1 时被破坏掉，我们应该选取 H_1 具有如下形状

$$H_1 = \begin{bmatrix} 1 & 0 \\ 0 & \widetilde{H}_1 \end{bmatrix} \begin{matrix} 1 \\ n-1 \end{matrix} \qquad (2.17)$$
$$\qquad\quad 1 \qquad n-1$$

利用形如（2.17）的 Householder 变换对 A 进行相似变换即有

$$H_1 A H_1 = \begin{bmatrix} a_{11} & a_2^{\mathrm{T}} \widetilde{H}_1 \\ \widetilde{H}_1 a_1 & \widetilde{H}_1 A_{22} \widetilde{H}_1 \end{bmatrix}, \qquad (2.18)$$

其中 $a_1^{\mathrm{T}} = (a_{21}, a_{31}, \cdots, a_{n1})$，$a_2^{\mathrm{T}} = (a_{12}, a_{13}, \cdots, a_{1n})$，$A_{22}$ 是 A 的右下角的 $n-1$ 阶主子阵。由（2.18）易知，Householder 变换 $\widetilde{H}_1$ 的最佳选择应该使得

$$\widetilde{H}_1 a_1 = p e_1, \qquad (2.19)$$

其中 $p \in \mathbb{R}$，e_1 是 $n-1$ 阶单位矩阵的第一列。这样一来，即可选取形如（2.17）的 Householder 变换 H_1，使（2.18）的第一列有 $n-2$ 个零元素。

然后，再对 $\widetilde{A}_{22} = \widetilde{H}_1 A_{22} \widetilde{H}_1$ 进行同样的考虑，又可找到 Householder 变换

$$\widetilde{H}_2 = \begin{bmatrix} 1 & 0 \\ 0 & \widetilde{\widetilde{H}}_2 \end{bmatrix} \in \mathbb{R}^{(n-1)\times(n-1)},$$

使得

$$(\widetilde{H}_2 \widetilde{A}_{22} \widetilde{H}_2) e_1 = \begin{bmatrix} * \\ * \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

从而令

$$H_2 = \begin{bmatrix} 1 & 0 \\ 0 & \widetilde{H}_2 \end{bmatrix},$$

即有

$$H_2H_1AH_1H_2=\begin{bmatrix} * & * & \\ * & * & * \\ 0 & * & \\ & & \ddots \\ 0 & & * \end{bmatrix}.$$

如此进行$n-2$步,即可找到$n-2$个 Householder 变换$H_1,\cdots,H_{n-2}$,使得

$$H_{n-2}\cdots H_1AH_1\cdots H_{n-2}=H,$$

其中 $H=[h_{ij}]$满足

$$h_{ij}=0,\qquad i>j+1,$$

即

$$H=\begin{bmatrix} h_1 & h_{12} & h_{13} & \cdots & h_{1,n-1} & h_{1n} \\ h_{21} & h_{22} & h_{23} & \cdots & h_{2,n-1} & h_{2n} \\ 0 & h_{32} & h_{33} & \cdots & h_{3,n-1} & h_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & h_{n,n-1} & h_{nn} \end{bmatrix}. \qquad (2.20)$$

通常称形如（2.20）的矩阵为**上 Hessenberg矩阵**。

现令

$$Q_0=H_1H_2\cdots H_{n-2}.$$

则有

$$Q_0^{\mathrm{T}}AQ_0=H. \qquad (2.21)$$

即上述过程将A正交相似变换成一个上 Hessenberg 矩阵 H（具有$\frac{1}{2}(n-1)(n-2)$个零元素）。通常称分解式（2.21）为A的**上 Hessenberg 分解**。

现在，我们再来考察对上 Hessenberg 矩阵H进行一次 QR 迭代

$$H=QR,\qquad \widetilde{H}=RQ$$

所需的运算量是多少。基于H的特殊形状，H的 QR 分解可用

Givens 变换实现，即计算$n-1$个 Givens 变换$G_i=G(i,i+1,\theta_i)$，$i=1,2,\cdots,n-1$，使

$$G_{n-1}G_{n-2}\cdots G_1A=R$$

为上三角矩阵；然后，$\widetilde{H}$ 可按如下方式计算

$$\widetilde{H}=RG_1^{\mathrm{T}}G_2^{\mathrm{T}}\cdots G_{n-1}^{\mathrm{T}}.$$

容易算出完成这一过程所需的运算量是$4n^2$(减少了一个数量级).

此外，容易证明这样得到的$\widetilde{H}$仍然是上 Hessenberg 矩阵。这样，我们就可一直迭代下去，每次迭代的运算量都是$4n^2$。

总结上面的讨论可知，先计算正交矩阵 Q_0，使 $Q_0^{\mathrm{T}}AQ_0=H$为上 Hessenberg 矩阵，然后再对H进行 QR 迭代，就可使每次迭代的运算量大为减少。

利用前面 Householder 变换约化 A 为上 Hessenberg 形的方法，可总结为如下的实用算法。

算法2.1

(1) 输入 $A=[a_{ij}]$，$k:=1$。

(2) 计算 $n-k$ 阶 Householder 变换 $\widetilde{H}_k$，使

$$\widetilde{H}_k\begin{bmatrix}a_{k+1,k}\\a_{k+2,k}\\\vdots\\a_{nk}\end{bmatrix}=\begin{bmatrix}*\\0\\\vdots\\0\end{bmatrix},$$

$$A:=H_kAH_k,$$

其中 $H_k=\mathrm{diag}(I_k,\widetilde{H}_k)$。

(3) 如果 $k<n-2$，则 $k:=k+1$ 转步(2)；否则，输出有关信息，结束。

这一算法计算出A的上 Hessenberg 形就存放在A所对应的存储单元内，所需运算量是$5n^3/3$；如果需要累积$Q_0=H_1H_2\cdots H_{n-1}$，则还需再增加运算量$2n^3/3$。

此外，用算法2.2计算得到的上 Hessenberg 阵$\hat{H}$满足

$$\hat{H}=Q^{\mathrm{T}}(A+E)Q,$$

其中Q是正交矩阵，$\|E\|_F\leqslant ch^2\|A\|_F\varepsilon$，这里$c$是一常数（详见文献[71]）。

当然，我们亦可用 Givens 变换将A约化为上 Hessenberg 形，一般需要的运算量为$\frac{10}{3}n^3$。但是，如果A有较多的零元素，则适当安排 Givens 变换的次序，可使运算量大为减少。另外，为了节省运算量，也可采用列主元的 Gauss 消去法，将A利用非正交的相似变换约化为上 Hessenberg 形。不过这样做，虽然运算量少，但数值稳定性较差。

尽管，一般来讲，上 Hessenberg 分解是不唯一的，然而，可以证明

定理2.2 设$A\in\mathbb{R}^{n\times n}$有如下两个上 Hessenberg 分解：

$$U^{\mathrm{T}}AU=H,\qquad V^{\mathrm{T}}AV=G, \tag{2.22}$$

其中$U=[u_1,\cdots,u_n]$和$V=[v_1,\cdots,v_n]$是n阶正交矩阵，$H=[h_{ij}]$和$G=[g_{ij}]$是上 Hessenberg 矩阵。若$u_1=v_1$，而且H的次对角元素$h_{i+1,i}$均不为零，则存在对角元素均为1或-1的对角矩阵D，使得

$$U=VD,\qquad H=DGD \tag{2.23}$$

（即$u_i=\pm v_i$，$|h_{ij}|=|g_{ij}|$，$i,j=1,2,\cdots,n$）。

证明 假定对某个m（$1\leqslant m<n$）已证：

$$u_j=\varepsilon_j v_j,\qquad j=1,2,\cdots,m, \tag{2.24}$$

其中$\varepsilon_1=1$，$\varepsilon_j=1$或-1。下面我们来证存在ε_{m+1}为-1或1使得

$$u_{m+1}=\varepsilon_{m+1}v_{m+1}.$$

从（2.22）可得

$$AU=UH,\qquad AV=VG.$$

分别比较上面两个矩阵等式的第m列，可得

$$Au_m = h_{1m}u_1 + \cdots + h_{mm}u_m + h_{m+1,m}u_{m+1}, \tag{2.25}$$

$$Av_m = g_{1m}v_1 + \cdots + g_{mm}v_m + g_{m+1,m}v_{m+1}, \tag{2.26}$$

分别在（2.25）和（2.26）两边左乘 u_i^T 和 $v_i^T(i=1,2,\cdots,m)$ 可得

$$h_{im} = u_i^T A u_m, \quad i=1,2,\cdots,m. \tag{2.27}$$

$$g_{im} = v_i^T A v_m, \tag{2.28}$$

由（2.24),(2.27）和（2.28）可得

$$h_{im} = \varepsilon_i \varepsilon_m g_{im}, \quad i=1,2,\cdots,m. \tag{2.29}$$

将（2.29）代入（2.25），并利用（2.24）和（2.26），可得

$$\begin{aligned} h_{m+1,m}u_{m+1} &= \varepsilon_m(Av_m - \varepsilon_1^2 g_{1m}v_1 - \cdots - \varepsilon_m^2 g_{mm}v_m) \\ &= \varepsilon_m(Av_m - g_{1m}v_1 - \cdots - g_{mm}v_m) \\ &= \varepsilon_m g_{m+1,m}v_{m+1}. \end{aligned} \tag{2.30}$$

由（2.30）即知

$$|h_{m+1,m}| = |g_{m+1,m}|,$$

而 $h_{m+1,m} \neq 0$，故（2.30）蕴含着

$$u_{m+1} = \varepsilon_{m+1} v_{m+1},$$

其中 $\varepsilon_{m+1} = 1$或-1。

因此，利用归纳法原理即知（2.23）成立。

一个上 Hessenberg 矩阵 $H=[h_{ij}]$，如果其次对角元素均不为零，即$h_{i+1,i} \neq 0$，$i=1,2,\cdots,n-1$，则称它是**不可约的**。定理2.2表明：如果 $Q^TAQ=H$ 为不可约的上 Hessenberg 矩阵，则 Q 和H完全由Q的第一列确定（这里是在相差一个正负号的意义下的唯一）。

2.4 双重步位移的 QR 迭代

下面我们来考虑如何加快 QR 迭代的收敛速度。假定A的特征值是$|\lambda_1| \geqslant \cdots \geqslant |\lambda_{n-1}| > |\lambda_n|$。前面关于正交迭代法的收敛性分析表明，随着 QR 迭代的进行，A_k 的右下角的对角元素$a_{nn}^{(k)}$将

收敛到特征值 λ_n，收敛速度是线性的，速率是$\left|\frac{\lambda_n}{\lambda_{n-1}}\right|$。由此可见，如果对某一常数 μ，将A的特征值重新编号，使其满足

$$|\lambda_1-\mu|\geqslant\cdots\geqslant|\lambda_{n-1}-\mu|>|\lambda_n-\mu|,$$

而且比值$|\lambda_n-\mu|/|\lambda_{n-1}-\mu|$很小，则用 $A-\mu I$ 代替A进行QR迭代可望其收敛速度更快。因此，这就促使我们考虑如下的迭代：

$$\begin{cases}H_1=Q_0^{\mathrm{T}}AQ_0 & (\text{上 Hessenberg 分解}),\\ H_k-\mu I=Q_kR_k & (\text{QR 分解}),\\ H_{k+1}=R_kQ_k+\mu I, & k=1,2,\cdots.\end{cases} \tag{2.31}$$

通常称迭代（2.31）为带位移的 QR 迭代，其中 μ 被称作**位移**。易见，这一迭代产生的 H_k 与A仍是正交相似的。

不失一般性，我们可以假定在迭代（2.31）中出现的上Hessenberg 矩阵都是不可约的。因若不然，在迭代的某一步，已有

$$H_k=\begin{bmatrix}H_{11}^{(k)} & 0\\ 0 & H_{22}^{(k)}\end{bmatrix},$$

则我们可以分别对 $H_{11}^{(k)}$ 和 $H_{22}^{(k)}$ 进行 QR 迭代即可。

现在我们来考虑如何有效地选取 μ。当然，大家从前面的分析已经明白，μ 选取的与A 的某个特征值越靠近越好。可是，问题是，在我们对A的特征值的信息知道甚少的前提下，如何选取 μ 才能使其与A的某个特征值较为靠近呢？

理论分析和实际计算的经验表明：QR 迭代产生的矩阵序列右下角最先显露A的特征值。因此，我们可以利用 QR 迭代的这一特点来选取位移 μ。如果显露的是 A 的实特征值，即 $h_{nn}^{(k)}$ 是 A 的较好的近似特征值时，当然，此时就可简单地选取$\mu=h_{nn}^{(k)}$作为第 $k+1$ 次迭代的位移。然而，如果显露的是A 的复共轭特征值时，即

$$G_k=\begin{bmatrix}h_{mm}^{(k)} & h_{mn}^{(k)}\\ h_{nm}^{(k)} & h_{nn}^{(k)}\end{bmatrix},\qquad m=n-1 \tag{2.32}$$

的特征值是一对互相共轭的复数 μ_1 和 μ_2，且与 A 的特征值比较接近时，就应该选择 G_k 的某一特征值 μ_i 作为位移。但这样一来就引进了复运算，而这是我们所不希望的。为了避免复运算的出现，人们想到用 μ_1 和 μ_2 连续作两次位移，即进行

$$H-\mu_1 I=U_1R_1,$$
$$H_1=R_1U_1+\mu_1 I,$$
$$H_1-\mu_2 I=U_2R_2,$$
$$H_2=R_2U_2+\mu_2 I,$$

这里我们记 $H=H_k$。对上面迭代产生的矩阵进行一些简单的推算，可得

$$M=QR, \tag{2.33}$$
$$H_2=Q^*HQ, \tag{2.34}$$

其中

$$M=(H-\mu_2 I)(H-\mu_1 I), \tag{2.35}$$
$$Q=U_1U_2,\qquad R=R_2R_1. \tag{2.36}$$

由(2.35)可得

$$M=H^2-sH+tI, \tag{2.37}$$

其中

$$s=\mu_1+\mu_2=h_{mm}^{(k)}+h_{nn}^{(k)}\in\mathbb{R},$$
$$t=\mu_1\mu_2=\det(G_k)\in\mathbb{R}.$$

因此 M 是一个实矩阵。而且如果在迭代过程中选取 R_1 和 R_2 的对角元素均为实数，则由(2.33)可推知，Q 亦是实的；从而由(2.34)知 H_2 亦是实的。这也就是说，在没有误差的情况下，用 μ_1 和 μ_2 连续作两次位移进行QR迭代产生的 H_2 仍是实的上Hessenberg 矩阵。但是，在实际计算时，由于舍入误差的影响，如此计算得到的 H_2 一般并不一定是实的。

因此，为了确保计算得到的 H_2 仍为实矩阵，根据(2.33)和(2.34)，我们自然想到按如下的步骤来计算 H_2：

(1) 计算 $M=H^2-sH+tI$；

(2) 计算M的 QR 分解：$M=QR$；

(3) 计算 $H_2=Q^{\mathrm{T}}HQ$。

然而，如此计算的第一步形成M就需运算量为$O(n^3)$，这使我们前面为减少每次迭代所需运算量而所做的努力全部付之东流。

幸运的是，定理 2.2 告诉我们：不论采用什么样的方法去求正交矩阵$\widetilde{Q}$使 $\widetilde{Q}^{\mathrm{T}}H\widetilde{Q}=\widetilde{H}_2$ 是上 Hessenberg 矩阵，只要保证$\widetilde{Q}$的第一列与Q的第一列一样，则$\widetilde{H}_2$就与H_2本质上是一样的(所有元素的绝对值相等)。因此，我们可有很大的自由度去寻求更有效的方法来实现由H到H_2的变换。

首先，我们从(2.33)知，Q的第一列与M的第一列共线，即Qe_1由Me_1单位化得到。而由(3.37)容易算出

$$Me_1=(\xi_1,\xi_2,\xi_3,0,\cdots,0)^{\mathrm{T}},$$

其中

$$\xi_1=(h_{11}^{(k)})^2+h_{12}^{(k)}h_{21}^{(k)}-sh_{11}^{(k)}+t,$$
$$\xi_2=h_{21}^{(k)}(h_{11}^{(k)}+h_{22}^{(k)}-s),\qquad \xi_3=h_{21}^{(k)}h_{32}^{(k)}.$$

其次，如果Householder 变换 P_0 将 Me_1 变为αe_1(即 $P_0(Me_1)=\alpha e_1$)，其中 $\alpha\in\mathbb{R}$，则易知，P_0 的第一列就与 Me_1 共线，从而 $P_0e_1=Qe_1$。而由第二章关于 Householder 变换的理论知，P_0可以按如下方式确定：

$$P_0=\operatorname{diag}(\widetilde{P}_0,I_{n-3}),$$

其中

$$\widetilde{P}_0=I_3-\beta vv^{\mathrm{T}},\qquad \beta=2/v^{\mathrm{T}}v,$$
$$v=(\xi_1-\alpha\,\mathrm{sign}(\xi_1),\xi_2,\xi_3)^{\mathrm{T}},$$
$$\alpha=(\xi_1^2+\xi_2^2+\xi_3^2)^{1/2}.$$

现令

$$B=P_0HP_0,$$

则我们只要能够找到第一列为 e_1 的正交矩阵$\widetilde{Q}$使 $\widetilde{Q}^{\mathrm{T}}B\widetilde{Q}=\widetilde{H}$ 为上 Hessenberg 矩阵，那么$\widetilde{H}$就是我们希望得到的H_2。由前面所介绍

的约化一个矩阵为上 Hessenberg 形的 Householder 方法可知，这是容易办到的。这只需确定$n-1$个Householder 变换$P_1,\cdots,P_{n-1}$，使

$$(P_{n-1}\cdots P_1)B(P_1\cdots P_{n-1})=\tilde{H}$$

为上Hessenberg矩阵，即有 $\tilde{Q}=P_1\cdots P_{n-1}$ 的第一列为e_1。而且由于B所具有的特殊性，实现这一约化过程所需的运算量仅为$O(n^2)$。

事实上，由于用P_0将H相似变换为B只改变了H的前三行前三列，故B具有如下形状

$$B=P_0HP_0=\begin{bmatrix} \times & \times & \times & \cdots & \times & \times \\ \times & \times & \times & \cdots & \times & \times \\ \oplus & \times & \times & \cdots & \times & \times \\ \oplus & \oplus & \times & \cdots & \times & \times \\ & & & \ddots & \vdots & \vdots \\ & 0 & & & & \\ & & & & \times & \times \end{bmatrix},$$

仅比上 Hessenberg 形多三个可能的非零元"$\oplus$"。由B的这种特殊性，易知用来约化B为上Hessenberg 形的第一个Householder变换P_1具有如下形状

$$P_1=\mathrm{diag}(1,\tilde{P}_1,I_{n-4}),$$

其中$\tilde{P}_1$为3 阶 Householder 变换，而且 P_1BP_1 具有如下形状

$$P_1BP_1=\begin{bmatrix} \times & \times & \times & \times & \cdots & \times & \times \\ \times & \times & \times & \times & \cdots & \times & \times \\ 0 & \times & \times & \times & \cdots & \times & \times \\ 0 & \oplus & \times & \times & \cdots & \times & \times \\ 0 & \oplus & \oplus & \times & \cdots & \times & \times \\ & & & & \ddots & \vdots & \vdots \\ & 0 & & & & & \\ & & & & & \times & \times \end{bmatrix}.$$

如此递推地进行，不难推出，第k次约化所用的Householder变换P_k具有如下形状

$$P_k=\mathrm{diag}(I_k,\tilde{P}_k,I_{n-k-3}),$$

其中 $\tilde{P}_k$ 为3阶 Householder 变换，$k=2,\cdots,n-3$，而且 $P_{n-3}\cdots P_1BP_1\cdots P_{n-3}$ 具有如下形状

$$P_{n-3}\cdots P_1BP_1\cdots P_{n-3}=\begin{bmatrix}\times & \cdots & \times & \times & \times\\ \times & \cdots & \times & \times & \times\\ & \ddots & \vdots & \vdots & \vdots\\ & 0 & \times & \times & \times\\ & & \oplus & \times & \times\end{bmatrix}.$$

因此，最后一次约化所用的 Householder 变换 P_{n-2}具有如下形状

$$P_{n-2}=\mathrm{diag}(I_{n-2},\tilde{P}_{n-2}),$$

其中$\tilde{P}_{n-2}$为2阶 Householder 变换。

这样我们就找到了一种实现由H到H_2的变换方法，它既避免了复运算的出现，又减少了运算量。当然，这一变换过程对G_k的两个特征值均为实数的情形亦是可行的。因此，我们就不必在选取位移时区别显露的是实特征值还是复共轭特征值的情形，而只需取作 G_k 的两个特征值即可。

综述上面的讨论，就得到了著名的 Francis 双重步位移的QR迭代算法：

算法2.2

(1) 输入不可约上 Hessenberg 矩阵$H=[h_{ij}]\in\mathbb{R}^{n\times n}$;

(2) $m:=n-1$，$k:=0$，$s:=h_{mm}+h_{nn}$,

$t:=h_{mm}h_{nn}-h_{mn}h_{nm}$，$x:=h_{11}^2+h_{12}h_{21}-sh_{11}+t$,

$y:=h_{21}(h_{11}+h_{22}-s)$，$z:=h_{21}h_{32}$;

(3) 如果 $k=n-2$，则转步(5)；否则，确定 Householder 矩阵 $\tilde{P}_k\in\mathbb{R}^{3\times3}$，使

$$\tilde{P}_k\begin{bmatrix}x\\ y\\ z\end{bmatrix}=\begin{bmatrix}*\\ 0\\ 0\end{bmatrix},$$

$H:=P_kHP_k(P_k=\text{diag}(I_k,\tilde{P}_k,I_{n-k-3}))$;

(4) $x:=h_{k+2,k+1}$, $y:=h_{k+3,k+1}$,

$$z:=\begin{cases} h_{k+4,k+1}, & \text{当 } k<n-3 \text{ 时}, \\ 0, & \text{当 } k=n-3 \text{ 时}, \end{cases}$$

$k:=k+1$，转步(3)；

(5) 确定 Householder 变换 $\tilde{P}_{n-2}\in\mathbb{R}^{2\times2}$，使

$$\tilde{P}_{n-2}\begin{bmatrix} x \\ y \end{bmatrix}=\begin{bmatrix} * \\ 0 \end{bmatrix},$$

$H:=P_{n-2}HP_{n-2}(P_{n-2}=\text{diag}(I_{n-2},\tilde{P}_{n-2}))$;

(6) 迭代结束.

这一算法的运算量是 $6n^2$；如果需要把正交变换累积起来，还需再增加运算量 $6n^2$.

2.5 双重步位移的QR算法

前面的讨论已解决了用QR方法求一个给定的实矩阵的实Schur分解的几个关键的问题. 然而，作为一种实用的算法，还需给出一种有效的判定准则，来判定迭代过程中所产生的上Hessenberg矩阵的次对角元素何时可以忽略不计. 一种简单而且适用的准则是：当

$$|h_{i+1,i}|\leqslant(|h_{ii}|+|h_{i+1,i+1}|)\varepsilon \tag{2.38}$$

时，就将 $h_{i+1,i}$ 看作是0. 这样做的理由是，在前面约化 A 为上Hessenberg矩阵 H 时就已引进了量级为 $\|A\|_F\varepsilon$ 的误差.

将算法2.1和2.2与收敛准则(2.38)结合起来，就得到了现在流行的双重步位移的QR算法，这一算法是计算一给定的 n 阶实矩阵 A 的实Schur分解：$Q^TAQ=T$，其中 Q 为正交矩阵，T 为拟上三角矩阵(即对角块为 1×1 或 2×2 方阵的块上三角矩阵).

算法2.3(QR算法)

(1) 输入 A.

(2) 上 Hessenberg 化：用算法2.1计算A的上 Hessenberg 分解 $H=U_0^{\mathrm{T}}AU_0$；$Q:=U_0$。

(3) 收敛性判定：

(i) 把所有满足条件

$$|h_{i,i-1}|\leqslant(|h_{ii}|+|h_{i-1,i-1}|)\varepsilon$$

的 $h_{i,i-1}$ 置零；

(ii) 确定最大的非负整数 m 和最小的非负整数 l，使

$$H=\begin{bmatrix} H_{11} & H_{12} & H_{13} \\ 0 & H_{22} & H_{23} \\ 0 & 0 & H_{33} \end{bmatrix}\begin{matrix} l \\ n-l-m \\ m \end{matrix},$$

$$\begin{matrix} l & n-l-m & m \end{matrix}$$

其中H_{33}为拟上三角形，而 H_{22} 为不可约的上 Hessenberg 形；

(iii) 如果 $m=n$，则输出有关信息，结束；否则进行下一步。

(4) 双重步位移QR迭代：对H_{22}用算法2.2迭代一次得$H_{22}:=P^{\mathrm{T}}H_{22}P$，其中$P=P_0P_1\cdots P_{n-m-l-2}$。

$$Q:=Q\,\mathrm{diag}(I_l,P,I_m),$$

$$H_{12}:=H_{12}P,\quad H_{23}:=P^{\mathrm{T}}H_{23}.$$

(5) 转步(3)。

实际计算的统计表明，这一算法每分离出一个 1 阶或 2 阶子矩阵，平均约需 2 次 Francis 迭 代。因此，平均来讲，如果只计算特征值，则这一算法的运算量大约是$8n^3$；如果Q和T都需要，则为$15n^3$。

误差分析的结果表明，这一算法计算所得到的实 Schur 标准形$\hat{T}$ 正交相似于一个非常靠近A的矩阵，即

$$Q^{\mathrm{T}}(A+E)Q=\hat{T},$$

其中$Q^{\mathrm{T}}Q=I$，$\|E\|_2\approx\|A\|_2\varepsilon$；计算所得到的 $\hat{Q}$ 是几乎正交的，即

$$\hat{Q}^{\mathrm{T}}\hat{Q}=I+F,\quad \|F\|_2\approx\varepsilon,$$

其中 ε 为机器精度，详见文献[37]。

§3 特征向量和不变子空间的计算

这一节，我们来讨论在用 QR 方法求得给定的矩阵的特征值之后，如何求其对应的特征向量和对应的不变子空间的一组标准正交基。

3.1 特征向量的计算

设$A\in\mathbf{R}^{n\times n}$，并假定已经利用QR方法求得$A$的特征值$\lambda$的一个近似值$\tilde{\lambda}$。现在，我们来讨论如何求$A$之对应于$\lambda$的近似特征向量。

目前，解决这一问题最好方法就是**反幂法**，它是乘幂法的自然推广(即将乘幂法用于A^{-1}上而得到的)，常用的是带位移的反幂法，其基本迭代格式如下：

$$(A-\mu I)v_k=z_{k-1}, \tag{3.1a}$$

$$z_k=v_k/\|v_k\|_2, \qquad k=0,1,2,\cdots, \tag{3.1b}$$

其中μ是事先选定的常数，称作**位移**；z_0是事先给定的向量，称作**初始向量**。

从(3.1)可以看出，每迭代一次就需解一个线性方程组，这要比乘幂法运算量大得多。但是，由于方程组的系数矩阵不随k的变化而变化，所以能够事先对它进行列选主元素的 LU 分解，然后每次迭代就只需解两个三角形方程组即可。

另外需要顺便指出的是，这里只是为了下面的分析方便，而在(3.1b)中用$\|\cdot\|_2$进行了规范化，在实际使用时是以$\|\cdot\|_\infty$进行规范化的。

假定A是非亏损的，即存在$X=[x_1,\cdots,x_n]\in\mathbf{C}^{n\times n}$非奇异，使得

$$X^{-1}AX=\Lambda\equiv\mathrm{diag}(\lambda_1,\cdots,\lambda_n), \tag{3.2}$$

而且还可以假定$\|x_i\|=1$，$i=1,2,\cdots,n$。现将初始向量z_0按x_1，$\cdots,x_n$展开：

$$z_0=\sum_{i=1}^{n}\beta_i x_i. \tag{3.3}$$

再假定μ与A的特征值λ_j最靠近，并有

$$0<|\mu-\lambda_j|<|\mu-\lambda_i|,\qquad i\neq j, \tag{3.4a}$$

$$\beta_j\neq 0. \tag{3.4b}$$

则从(3.1a)，(3.2)和(3.3)可得

$$\begin{aligned} v_k &= \theta_k(A-\mu I)^{-k}z_0 \\ &= \theta_k\sum_{i=1}^{n}\beta_i(\lambda_i-\mu)^{-k}x_i \\ &= \theta_k\beta_j(\lambda_j-\mu)^{-k}(x_j+u_k), \end{aligned} \tag{3.5}$$

其中θ_k是一个正数，而

$$u_k=\sum_{i\neq j}\left(\frac{\lambda_j-\mu}{\lambda_i-\mu}\right)^k\frac{\beta_i}{\beta_j}x_i\to 0,\qquad k\to\infty,$$

其收敛速度依赖于$|\lambda_j-\mu|/\min\limits_{i\neq j}|\mu-\lambda_i|$的大小。将(3.5)代入(3.1b)即得

$$z_k=\frac{v_k}{\|v_k\|_2}=\frac{\eta_k}{\|x_j+u_k\|_2}(x_j+u_k),$$

其中η_k为满足$|\eta_k|=1$的复数。从而有

$$\begin{aligned} &\operatorname{dist}(\mathscr{R}(z_k),\mathscr{R}(x_j)) \\ &\quad=\|z_kz_k^*-x_jx_j^*\|_2 \\ &\quad=\left\|\frac{(x_j+u_k)(x_j+u_k)^*}{(x_j+u_k)^*(x_j+u_k)}-x_jx_j^*\right\|_2 \\ &\quad\to 0,\qquad k\to\infty, \end{aligned}$$

收敛速度依赖于$|\lambda_j-\mu|/\min\limits_{i\neq j}|\mu-\lambda_i|$的大小。换句话说，就是$z_k$将按方向收敛于$A$的特征向量，$\mu$与$\lambda_j$越靠近，收敛速度就越

快。

由此可见，从收敛速度的角度来考虑，用(3.1)进行迭代时，μ取得越靠近A的某个特征值越好。但是，当μ与A的特征值很靠近时，$A-\mu I$就与一个奇异矩阵很靠近，每迭代一步就需解一个非常病态的线性方程组。然而，实际计算的经验和理论分析的结果表明：$A-\mu I$的病态性，并不影响其收敛速度，而且当μ与A的某个特征值很靠近时，常常只需迭代一次就可得到相当好的近似特征向量。为了弄清这点，我们作下面的简要分析。

首先，我们来确立一个判定一个向量v是否可作为A的近似特征向量的判定准则。设λ是A的一个特征值，v是$\mathbb{C}^n$中的一个单位向量($\|v\|_2=1$)，定义

$$r=(A-\lambda I)v \tag{3.6}$$

为向量v的剩余向量。从(3.6)易得

$$(A-rv^*)v=\lambda v,$$

即v是$A-rv^*$对应于λ的特征向量。如果$\|r\|_2$很小，则$\|rv^*\|_2$亦很小；从而如果A之对应于λ的特征向量是良态的话，当$\|r\|_2$很小时，v就是A之对应于λ的一个很好的近似特征向量。因此，我们可用v的剩余向量的大小来恒量v是否可作为对应于λ的近似特征向量。

现在，假定

$$|\mu-\lambda|=\min_{\tilde{\lambda}\in\lambda(A)}|\tilde{\lambda}-\mu|\leqslant\varepsilon_1, \tag{3.7}$$

其中ε_1是一个很小的正数(通常有$\varepsilon_1=O(\varepsilon)$)。再假定对给定的初始向量$z_0$是用列主元素的Gauss消去法求解方程组(3.1a)得到向量v_1的计算值$\tilde{v}_1$的。则由Gauss消去法的误差分析结果知

$$(A-\mu I+E)\tilde{v}_1=z_0, \tag{3.8}$$

其中$\|E\|_2\leqslant\varepsilon_2$(通常$\varepsilon_2=O(\varepsilon)$)。这样，由(3.1b)计算得

$$z_1=\tilde{v}_1/\|\tilde{v}_1\|_2, \tag{3.9}$$

这里为了避免符号上的麻烦，忽略了计算z_1所引起的误差。

由(3.8)可得向量 z_1 的剩余向量为

$$r=(A-\lambda I)z_1=(\mu-\lambda)z_1-Ez_1+z_0/\|\tilde{v}_1\|_2.$$

于是

$$\|r\|_2\leqslant\varepsilon_1+\varepsilon_2+\|\tilde{v}_1\|_2^{-1}, \tag{3.10}$$

这里假定 $\|z_0\|_2=1$。

由此可见，如果计算得到的 $\tilde{v}_1$ 有很大的范数，则由(3.1)迭代一次所得到的向量就有范数很小的剩余向量，从而在特征值问题不是十分病态的条件下，就得到了很好的近似特征向量。

现在来说明在条件(3.7)成立的前提下确有 $\tilde{v}_1$ 的范数很大。

设 $A-\mu I+E$ 的奇异值分解为

$$A-\mu I+E=U\Sigma W^*, \tag{3.11}$$

其中 $U=[u_1,\cdots,u_n]$，$W=[w_1,\cdots,w_n]\in\mathscr{U}_n$，$\Sigma=\mathrm{diag}(\sigma_1,\cdots,\sigma_n)$，$\sigma_1\geqslant\cdots\geqslant\sigma_n>0$。

由 $\lambda-\mu$ 是 $A-\mu I$ 的特征值易得

$$\sigma_n(A-\mu I)\leqslant|\lambda-\mu|\leqslant\varepsilon_1,$$

其中 $\sigma_n(A-\mu I)$ 表示 $A-\mu I$ 的最小奇异值；再由第一章的推论6.6得

$$\sigma_n\leqslant\sigma_n(A-\mu I)+\|E\|_2\leqslant\varepsilon_1+\varepsilon_2. \tag{3.12}$$

将 z_0 按 $u_1,\cdots,u_n$ 展开，有

$$z_0=\sum_{i=1}^n\alpha_ju_j,\qquad\sum_{j=1}^n|\alpha_j|^2=\|z_0\|_2=1.$$

从而有

$$\begin{aligned}\tilde{v}_1&=(A-\mu I+E)^{-1}z_0\\&=W\Sigma^{-1}U^*\Big(\sum_{j=1}^n\alpha_ju_j\Big)=\sum_{j=1}^n\frac{\alpha_j}{\sigma_j}w_j,\end{aligned} \tag{3.13}$$

于是

$$\|\tilde{v}_1\|_2=\Big(\sum_{j=1}^n\Big|\frac{\alpha_j}{\sigma_j}\Big|^2\Big)^{\frac{1}{2}}\geqslant\frac{|\alpha_n|}{\sigma_n}\geqslant\frac{|\alpha_n|}{\varepsilon_1+\varepsilon_2}.$$

这样一来，只要$|\alpha_n|$不是很小(即z_0在u_n方向上不是十分亏损)的话，则$\|\tilde{v}_1\|_2$就很大。因此，通常反幂法只需迭代一次就足够了。

这里还需指出的一点是，在λ比较病态时，利用反幂法再进行第二次迭代，一般不会得到更好的近似特征向量。这是由于$\tilde{v}_1$的单位化向量z_1可表示为

$$z_1 = \gamma_n w_n + \sum_{j=1}^{n-1} \gamma_j w_j,$$

其中γ_n非常接近1，$\gamma_j (j = 1,2,\cdots,n-1)$都是小量；第二次迭代时，将$z_1$按$u_1,\cdots,u_n$展开，其系数将主要由$\gamma_n w_n$决定，即

$$z_1 \approx (\gamma_n w_n^* u_n) u_n + \sum_{j=1}^{n} (\gamma_n w_n^* u_j) u_j;$$

而此时σ_n很小，因而u_n和w_n将与A对应于λ的左右特征向量很接近，从而$|u_n^* w_n| \approx s(\lambda)$将很小，即此时$z_1$在$u_n$方向上的投影几乎等于零。因此，第二次迭代得到的$z_2$就不会有范数很小的剩余向量。现在看一个具体的例子。

设

$$A = \begin{bmatrix} 1 & 1 \\ 10^{-10} & 1 \end{bmatrix}.$$

它有特征值$\lambda_1 = 0.99999$和$\lambda_2 = 1.00001$，以及对应的特征向量$x_1 = (1, -10^{-5})^T$和$x_2 = (1, 10^{-5})^T$。两个特征值的条件数均为10^5数量级。取$\mu = 1$，$z_0 = (0,1)^T$，应用反幂法迭代一次(在十位十进制的浮点数系下进行)，则$z_1 = (1,0)^T$，并且$\|Az_1 - \mu z_1\|_2 = 10^{-10}$。可是再迭代一次将产生$z_2 = (0,1)^T$，则有$\|Az_2 - \mu z_2\|_2 = 1$。

上述分析表明，利用反幂法求特征向量时，位移量μ取作较精确的近似特征值最好。此时，一般只需迭代一次就可得到很好的近似特征向量。因此，通常总是在用某种方法求得A的近似特征值之后，再利用反幂法求对应的特征向量。连同QR方法一起来使用反幂法的基本步骤如下：

（1）计算A的Hessenberg分解：$U_0^{\mathrm{T}}AU_0=H$；

（2）使用双重步位移的QR方法求出H的特征值，而不累积变换阵；

（3）对每个计算的特征值$\hat{\lambda}$，在(3.1)中取$A=H$，$\mu=\hat{\lambda}$进行迭代，求出向量z，使$Hz\approx\hat{\lambda}z$；

（4）计算$x=U_0z$(则x就是对应于$\hat{\lambda}$的近似特征向量)。

注3.1 在上述方法的第三步用反幂法迭代时，其收敛性是用向量$r=(A-\mu I)z$的大小来恒量的；迭代的初始向量，是采用随机选取的方法来选取的，在实际计算时，有专门实现随机选取的程序。

3.2 不变子空间的计算

设$A\in\mathbb{R}^{n\times n}$的特征值$\lambda_1,\cdots,\lambda_n$满足$\{\lambda_1,\cdots,\lambda_l\}\cap\{\lambda_{l+1},\cdots,\lambda_n\}=\varnothing$，并假定集$\mathscr{L}=\{\lambda_1,\cdots,\lambda_l\}$关于复共轭是封闭的(即若$\lambda\in\mathscr{L}$，则必有$\bar{\lambda}\in\mathscr{L}$)，则在$\mathbb{R}^n$中存在唯一的$l$维子空间$\mathscr{X}$使得$A\mathscr{X}\subset\mathscr{X}$，而且$A$在$\mathscr{X}$上的限制$A|_{\mathscr{X}}$的特征值正好是$\lambda_1,\cdots,\lambda_l$，即由$\lambda_1,\cdots,\lambda_l$可唯一确定$A$的一个$l$维不变子空间。下面我们将主要讨论如何求由$\lambda_1,\cdots,\lambda_l$确定的不变子空间$\mathscr{X}$的一组标准正交基。

大家知道，如果已知A的实Schur分解为

$$Q^{\mathrm{T}}AQ=\begin{bmatrix}T_{11} & T_{12}\\ 0 & T_{22}\end{bmatrix}\begin{matrix}l\\ n-l\end{matrix},\quad \begin{matrix}l & \ \ n-l\end{matrix}$$

而且T_{11}的特征值正好是$\lambda_1,\cdots,\lambda_l$，则$Q$的前$l$列就是对应于$\lambda_1,\cdots,\lambda_l$的$A$的$l$维不变子空间$\mathscr{X}$的一组标准正交基。遗憾的是，如果我们是用双重步位移的QR方法来实现A的实Schur分解的话，则所得到的拟上三角形的对角块是非常任意的，并不能按我们事先需要的次序排列，即若计算得到的Schur分解是

$$\hat{Q}^{\mathrm{T}}A\hat{Q}=\hat{T}\equiv\begin{bmatrix}\hat{T}_{11} & \hat{T}_{12}\\ 0 & \hat{T}_{22}\end{bmatrix}\begin{matrix}l\\ n-l\end{matrix},$$
$$\begin{matrix}\quad l & \quad n-l\end{matrix}$$

则我们并不能期望有$\lambda(\hat{T}_{11})=\mathscr{L}$成立．但是如果我们能够找到一个正交矩阵$\tilde{Q}$使得

$$\tilde{Q}^{\mathrm{T}}\hat{T}\tilde{Q}=\tilde{T}\equiv\begin{bmatrix}\tilde{T}_{11} & \tilde{T}_{12}\\ 0 & \tilde{T}_{22}\end{bmatrix}\begin{matrix}l\\ n-l\end{matrix}$$
$$\begin{matrix}\quad l & \quad n-l\end{matrix}$$

仍是拟上三角矩阵，且$\lambda(\tilde{T}_{11})=\mathscr{L}$，则问题也就解决了．这样一来，我们要求指定的$\lambda_1,\cdots,\lambda_l$所对应的$A$的不变子空间的一组标准正交基的问题，就归结为按指定次序来重排一个拟上三角矩阵的对角块的问题．

先来考虑A是2×2矩阵的情形．假设我们已计算出正交矩阵$\hat{Q}$使

$$\hat{Q}^{\mathrm{T}}A\hat{Q}=\hat{T}=\begin{bmatrix}\lambda_1 & t_{12}\\ 0 & \lambda_2\end{bmatrix},\quad \lambda_1\neq\lambda_2,$$

希望调换λ_1和λ_2的位置，即确定一个正交矩阵$\tilde{Q}$，使得

$$\tilde{Q}^{\mathrm{T}}\hat{T}\tilde{Q}=\begin{bmatrix}\lambda_2 & *\\ 0 & \lambda_1\end{bmatrix}.$$

容易算出$\hat{T}$对应于λ_2的特征向量是

$$x=(t_{12},\lambda_2-\lambda_1)^{\mathrm{T}},$$

即有$\hat{T}x=\lambda_2x$．再取正交矩阵$\tilde{Q}$为使$\tilde{Q}^{\mathrm{T}}x$的第二个分量为零的Givens变换，则

$$\tilde{Q}^{\mathrm{T}}\hat{T}\tilde{Q}e_1=\lambda_2e_1.$$

因此，令$Q=\hat{Q}\tilde{Q}$，则

$$Q^{\mathrm{T}}AQ=\begin{bmatrix}\lambda_2 & \pm t_{12}\\ 0 & \lambda_1\end{bmatrix}.$$

这样我们就解决了$n=2$时交换λ_1和λ_2的问题．

对于一般情形，在A的特征值全是实数的条件之下，使用上

述技巧有计划地交换相邻的特征值，就可把 $\lambda(A)$ 的任意子集移到 T 之对角线的前面，从而求得所需的不变子空间的一组标准正交基。具体算法如下。

算法3.1

(1) 输入上三角矩阵 $T=[t_{ij}]$，正交矩阵 Q 和指定的特征值子集 $\mathscr{L}=\{\lambda_1,\cdots,\lambda_l\}$。

(2) 如果 $\{t_{11},\cdots,t_{ll}\}=\mathscr{L}$，则输出有关信息结束；否则，$k:=1$，进行下一步。

(3) 若 $t_{kk}\notin\mathscr{L}$ 而 $t_{k+1,k+1}\in\mathscr{L}$，则确定 $c=\cos\theta$ $s=\sin\theta$，使

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix}\begin{bmatrix} t_{k,k+1} \\ t_{k+1,k+1}-t_{kk} \end{bmatrix}=\begin{bmatrix} * \\ 0 \end{bmatrix},$$

$$T:=G_kTG_k^{\mathrm{T}},\quad Q:=QG_k,(G_k=G(k,k+1,\theta));$$

否则，进行下一步。

(4) 如果 $k<n-1$，则 $k:=k+1$，转步(3)；否则转步(2)。

注3.2 其中输入的 T 和 Q 是由已知方阵 A 应用双重步位移的 QR 方法计算得到的，即

$$Q^{\mathrm{T}}AQ=T.$$

当 A 的特征值并非都是实数时，利用 QR 方法求得的实 Schur 标准形之对角块就必含有 2×2 的块，此时就会涉及到 2×2 块的交换问题。具体地讲，假定

$$T=\begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix},$$

其中 T_{11} 是1阶或2阶方阵，T_{22} 是具有一对复共轭特征值的2阶方阵，且 $\lambda(T_{11})\cap\lambda(T_{22})=\varnothing$，我们需要确定一个正交矩阵 Q，使得

$$Q^{\mathrm{T}}TQ=\begin{bmatrix} \hat{T}_{11} & \hat{T}_{12} \\ 0 & \hat{T}_{22} \end{bmatrix},$$

满足 $\hat{T}_{11}$ 是2阶方阵，且 $\lambda(\hat{T}_{11})=\lambda(T_{22})$。这可用前面类似的技

巧完成，不过要稍微复杂一些，请读者作为练习自己补出。

§4 对称 QR 方法

对称 QR 方法就是求解实对称特征值问题的 QR 方法，是将 QR方法应用于对称矩阵，并充分利用其对称性而得到的；是目前求解中小型稠密对称矩阵的特征值和对应的特征向量的最有效方法之一。

设 $A\in S\mathbb{R}^{n\times n}$。我们知道，为了减少每次 QR 迭代所需的运算量，QR 方法的第一步就是将给定的矩阵 A 约化为上 Hessenberg 矩阵，即计算正交矩阵 U_0，使得

$$U_0^{\mathrm{T}}AU_0=T \tag{4.1}$$

为上 Hessenberg 矩阵。而当 A 对称时，T 亦是对称的，从而 T 是对称三对角矩阵。因此，对实对称矩阵而言，我们首先应该将给定的矩阵 A 三对角化。而且，在约化过程中再充分利用对称性，还可使约化的运算量大为减少。

将 A 作如下分块

$$A=\begin{bmatrix}\alpha_1 & v_0^{\mathrm{T}}\\ v_0 & A_0\end{bmatrix}\begin{matrix}1\\ n-1\end{matrix},\quad \begin{matrix}1 & n-1\end{matrix}$$

从约化一个矩阵为上 Hessenberg 矩阵的 Householder 方法不难推出，利用 Householder 变换将 A 约化为对称三对角阵的第 k 步为：

(i) 计算 Householder 变换 $\widetilde{H}_k\in\mathbb{R}^{(n-k)\times(n-k)}$，使得

$$\widetilde{H}_k v_{k-1}=\beta_k e_1,\quad \beta_k\in\mathbb{R};$$

(ii) 计算

$$\begin{matrix}1\\ n-k-1\end{matrix}\begin{bmatrix}\alpha_{k+1} & v_k^{\mathrm{T}}\\ v_k & A_k\end{bmatrix}=\widetilde{H}_kA_{k-1}\widetilde{H}_k,\quad \begin{matrix}1 & n-k-1\end{matrix}$$

这里 $k=1,2,\cdots,n-1$。

如果用上述约化过程产生的 α_k,β_k 和 $\tilde{H}_k$ 定义

$$T=\begin{bmatrix}\alpha_1 & \beta_1 & & 0\\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{n-1}\\ 0 & & \beta_{n-1} & \alpha_n\end{bmatrix},\quad H_k=\begin{bmatrix}I_k & 0\\ 0 & \tilde{H}_k\end{bmatrix},$$

$$Q=H_1H_2\cdots H_{n-1},$$

则有

$$Q^{\mathrm{T}}AQ=T.$$

从上述约化过程容易看出，第 k 步约化的主要工作量是计算 $\tilde{H}_kA_{k-1}\tilde{H}_k$。设

$$\tilde{H}_k=I-\beta vv^{\mathrm{T}},\quad v\in\mathbb{R}^{n-k}.$$

则利用 A_{k-1} 的对称性，易得

$$\tilde{H}_kA_{k-1}\tilde{H}_k=A_{k-1}-vw^{\mathrm{T}}-w^{\mathrm{T}}v,\tag{4.2}$$

其中

$$w=u-\frac{1}{2}\beta(v^{T}u)v,\quad u=\beta A_{k-1}v.\tag{4.3}$$

利用（4.2）和（4.3），并注意到对称性，容易设计出运算量为 $2(n-k)^2$ 的计算$\tilde{H}_kA\tilde{H}_k$的算法。因此，完成整个约化所需的运算量为 $2n^3/3$，而非对称时把一个 n 阶方阵约化为上 Hessenberg 阵所需的运算量为 $5n^3/3$。

上述三对角化过程的详情细节，请读者作练习自己补出，并在此基础上总结出一个实用的算法。

在完成了把 A 约化为对称三对角矩阵 T 的任务之后，我们要做的第二件事就是选取适当的位移进行 QR 迭代。由于此时 A 的特征值全是实数，因而再使用双重步位移是完全没有必要的，只需进行单步位移即可。

设 T 是对称三对角矩阵。我们来考虑单步位移的 QR 迭代：

$$\begin{cases}T_k-\mu I=Q_kR_k,\\ T_{k+1}=R_kQ_k+\mu I,\end{cases}\quad k=0,1,2,\cdots,\tag{4.4}$$

其中 $T_0=T$，Q_k 是正交矩阵，R_k 是上三角矩阵。根据 QR 迭代保持上 Hessenberg 形和对称性不变的特点，我们立即知道迭代(4.4)产生的 T_k 都是对称三对角矩阵。此外，容易算出每次迭代只需 $O(n)$ 次运算即可完成，这比非对称的情形降低了一个数量级。

与非对称的 QR 迭代一样，这里我们亦可假定迭代所用的 H_k 均是不可约的，即

$$H_k=\begin{bmatrix} \alpha_1^{(k)} & \beta_1^{(k)} & & 0 \\ \beta_1^{(k)} & \ddots & \ddots & \\ & \ddots & \ddots & \beta_{n-1}^{(k)} \\ 0 & & \beta_{n-1}^{(k)} & \alpha_n^{(b)} \end{bmatrix}$$

满足 $\beta_i^{(k)}\neq 0$，$i=1,2,\cdots,n-1$。

现在再来看位移 μ 应该如何选取。从非对称 QR 方法的讨论中，我们知道一种简单易行的取法是取 $\mu=\alpha_n^{(k)}$（通常称作**Rayleigh 商位移**）。但是，另一种更好的选法就是取 μ 为矩阵

$$\begin{bmatrix} \alpha_{n-1}^{(k)} & \beta_{n-1}^{(k)} \\ \beta_{n-1}^{(k)} & \alpha_n^{(k)} \end{bmatrix},$$

靠近 $\alpha_n^{(k)}$ 的特征值，即取

$$\mu=\alpha_n^{(k)}+\delta-\operatorname{sign}\delta\cdot\sqrt{\delta^2+(\beta_{n-1}^{(k)})^2}, \tag{4.5}$$

其中 $\delta=\frac{1}{2}(\alpha_{n-1}^{(k)}-\alpha_n^{(k)})$。这就是著名的 **Wilkinson 位移**。Wilkinson（参见文献[71]）曾证明了这两种位移都是最终三次收敛的，并说明了为什么后者较前者更好一些。

下面再来考虑如何具体实现一步带位移的对称 QR 迭代：

$$\begin{cases} T-\mu I=QR, \\ \hat{T}=RQ+\mu I, \end{cases} \tag{4.6}$$

其中 Q 为正交矩阵，R 为上三角矩阵，

$$T=\begin{bmatrix} \alpha_1 & \beta_1 & & & 0 \\ \beta_1 & \alpha_2 & \ddots & & \\ & \beta_2 & \ddots & \ddots & \\ & & \ddots & \alpha_{n-1} & \beta_{n-1} \\ 0 & & & \beta_{n-1} & \alpha_n \end{bmatrix}, \quad \beta_i \neq 0.$$

当然，我们可以利用 Givens 变换来实现 $T-\mu I$ 的 QR 分解，进而完成一步迭代。但是，更漂亮的方法是利用定理 2.2 以隐含的方式来实现由 T 到 $\hat{T}$ 的变换。

大家知道，迭代(4.6)的实质是将 T 用正交相似变换变为 $\hat{T}$，即 $\hat{T}=Q^{\mathrm{T}}TQ$。因而从定理 2.2 知道 $\hat{T}$ 本质上由 Q 的第一列完全确定.从利用 Givens 变换实现 $T-\mu I$ 的 QR 分解过程易知 $Qe_1=G_1e_1$，其中 G_1 是将 $T-\mu I$ 的第一列之第二个元素变为零的 Givens 变换，即

$$G_1=G(1,2,\theta),$$

其中 θ 满足

$$\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \alpha_1-\mu \\ \beta_1 \end{bmatrix} = \begin{bmatrix} * \\ 0 \end{bmatrix}.$$

令

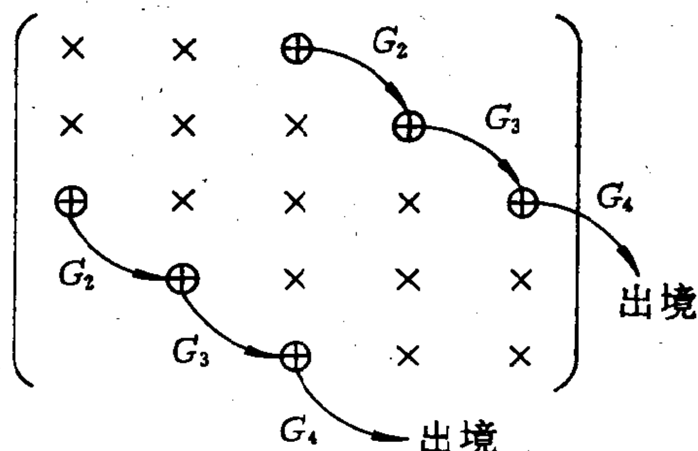
$$B=G_1^{\mathrm{T}}TG_1,$$

则 B（例如，$n=5$）有如下形状

$$B=\begin{bmatrix} \times & \times & \oplus & 0 & 0 \\ \times & \times & \times & 0 & 0 \\ \oplus & \times & \times & \times & 0 \\ 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & \times & \times \end{bmatrix},$$

仅比三对角矩阵多两个非零元“$\oplus$”.由此即知，只需用 Givens 变换将 B 约化为三对角矩阵即可得到所需的三对角矩阵 $\hat{T}$。下面就 $n=5$ 的情形，来说明这一约化过程。

首先用（2,3）坐标平面上的 Givens 变换 G_2 将 B 之(1,3)和(3,1)位置上的元素消为0，这样在(2,4)和(4,2)位置上又会出现两个我们不希望有的非零元；接着再用（3,4)平面上的 Givens 变换 G_3 将（2,4）和（4,2）位置上的元素消为零，又在（3,5）和（5,3）位置上出现两个非零元；最后利用（4,5）平面上的 Givens 变换 G_4 将（3,5）和（5,3）位置上的元素消为零，而得到所需的三对角矩阵。这一过程可用图示方式形象地表述如下：



从图示可以看出，整个约化过程就是把三对角线之外不受欢迎的非零元“⊕”逐步“驱赶”出矩阵之外。因此，我们可以称这一约化方法为“驱逐出境法”。

对于一般的 n，我们可以通过 $n-2$ 次 Givens 变换 $G_2,\cdots,G_{n-1}$ 使

$$G_{n-1}^{\mathrm{T}}\cdots G_2^{\mathrm{T}}(G_1^{\mathrm{T}}TG_1)G_2\cdots G_{n-1}$$

是三对角矩阵，且 $(G_1\cdots G_{n-1})e_1=G_1e_1=Qe_1$。

综上所述，可得带 Wilkinson 位移的对称 QR 迭代算法如下：

算法4.1

（1）输入 T 的对角元素 $\alpha_1,\cdots,\alpha_n$ 和次对角元素 $\beta_1,\cdots,\beta_{n-1}$，$k:=1$。

（2）$\delta:=(\alpha_{n-1}-\alpha_n)/2$，

$$\mu := \alpha_n - \beta_{n-1}^2 / (\delta + \mathrm{sign}\delta \cdot \sqrt{\delta^2 + \beta_{n-1}^2}),$$
$$x := \alpha_1 - \mu,$$
$$y := \beta_1.$$

（3）计算 $c = \cos\theta$，$s = \sin\theta$ 和 σ 使

$$\begin{bmatrix} c & -s \\ s & c \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \sigma \\ 0 \end{bmatrix},$$

$$\begin{bmatrix} \alpha_k & \beta_k \\ \beta_k & \alpha_{k+1} \end{bmatrix} := \begin{bmatrix} c & -s \\ s & c \end{bmatrix} \begin{bmatrix} \alpha_k & \beta_k \\ \beta_k & \alpha_{k+1} \end{bmatrix} \begin{bmatrix} c & s \\ -s & c \end{bmatrix},$$

$$\beta_{k+1} := c\beta_{k+1}.$$

（4）如果 $k>1$，则 $\beta_{k-1} := \sigma$；否则进行下一步。

（5）如果 $k<n-1$，则

$$x := \beta_k, \quad y := -s\beta_{k+1}, \quad k := k+1,$$

转步(3)；否则迭代结束。

这一算法运算量为：乘除运算 $20n$，开方 $n-1$ 次；如果对给定的正交矩阵 Q，还需计算 $QG_1 \cdots G_{n-1}$，则还需再增加运算量 $4n^2$。

至此，我们已经解决了对称 QR 方法的几个关键问题，可以总结为如下算法：

算法4.2（对称 QR 算法）

（1）输入 $A \in S\mathrm{R}^{n\times n}$ 和误差限 ε。

（2）三对角化 A：

计算 Householder 变换 $H_1, \cdots, H_{n-2}$，使

$$(H_1 \cdots H_{n-2})^{\mathrm{T}} A (H_1 \cdots H_{n-2})$$

$$= \begin{bmatrix} \alpha_1 & \beta_1 & & & 0 \\ \beta_1 & \alpha_2 & \ddots & & \\ & \beta_2 & \ddots & \ddots & \\ & & \ddots & \alpha_{n-1} & \beta_{n-1} \\ 0 & & & \beta_{n-1} & \alpha_n \end{bmatrix}.$$

(3) 收敛性检验：

(i) 将所有满足

$$|\beta_i| \leqslant \varepsilon(|\alpha_i| + |\alpha_{i+1}|)$$

的 β_i 置零；

(ii) 如果 $\beta_i = 0$，$i = 1,2,\cdots,n-1$，则输出有关信息，结束；否则 $\beta_0 := 0$，确定正整数 $p < q$，使得，

$$\beta_{p-1} = 0, \quad \beta_i \neq 0, \ i = p, p+1, \cdots, q,$$
$$\beta_{q+1} = \cdots = \beta_{n-1} = 0.$$

(4) QR 迭代：

对对称三对角矩阵

$$T = \begin{bmatrix} \alpha_p & \beta_p & & 0 \\ \beta_p & \alpha_{p+1} & \ddots & \\ & \ddots & \ddots & \beta_q \\ 0 & & \beta_q & \alpha_q \end{bmatrix}$$

应用算法 4.1，然后转(3)。

这一算法是数值代数中最漂亮的算法。误差分析的结果表明，由算法 4.2 计算所得到的特征值 $\hat{\lambda}_1, \cdots, \hat{\lambda}_n$ 满足

$$Q^{\mathrm{T}}(A+E)Q = \operatorname{diag}(\hat{\lambda}_1, \cdots, \hat{\lambda}_n),$$

其中 $Q^{\mathrm{T}}Q = I$，$\|E\|_2 \approx \|A\|_2 \boldsymbol{\varepsilon}$。从而由第一章的 Weyl 定理知，

$$|\hat{\lambda}_i - \lambda_i| \approx \|A\|_2 \boldsymbol{\varepsilon}, \quad i = 1, 2, \cdots, n,$$

其中 λ_i 是 A 的精确特征值，且与 $\hat{\lambda}_i$ 排列次序一致。这也就是说，对称 QR 算法计算得到的特征值是相当精确的，相对误差不超过机器精度。但值得注意的是，特征向量的计算值并不一定亦有这样的精度，它与 λ_i 和其他特征值的分离程度有关。

§5 奇异值分解的计算

设 $A \in \mathbb{R}^{m \times n}(m \geqslant n)$。从奇异值分解定理的证明过程可知，

A的奇异值分解可从实对称矩阵

$$C = A^{\mathrm{T}} A$$

的Schur分解导出。因此，我们自然想到先利用对称QR方法来实现C的Schur分解，然后借助C的Schur分解来实现A的奇异值分解。然而，这样做有两个缺点：一是计算$A^{\mathrm{T}}A$的运算量较大；二是计算$A^{\mathrm{T}}A$容易引入较大的误差，有时甚至使计算所得到的奇异值面貌全非。针对这样的问题，Golub和Kahan于1965年提出了一种十分稳定而有效的计算奇异值分解的方法。他们的基本思想就是隐含地应用对称QR方法于AA^{T}上，而并不需明确地将$A^{\mathrm{T}}A$计算出来，从而避免了上述两个问题的出现。

要应用对称QR方法于$A^{\mathrm{T}}A$上，第一步就需将其三对角化。这一步为了避免$A^{\mathrm{T}}A$的计算，可先将A二对角化，即求正交矩阵U_1和V_1，使

$$U_1^{\mathrm{T}} A V_1 = \begin{bmatrix} B \\ 0 \end{bmatrix} \begin{matrix} n \\ m-n \end{matrix}, \tag{5.1}$$

其中

$$B = \begin{bmatrix} \delta_1 & \gamma_2 & & & 0 \\ & \delta_2 & \gamma_3 & & \\ & & \ddots & \ddots & \\ & & & \ddots & \gamma_n \\ 0 & & & & \delta_n \end{bmatrix}. \tag{5.2}$$

事实上，一旦分解（5.1）已经实现，则有

$$V_1^{\mathrm{T}} A^{\mathrm{T}} A V_1 = B^{\mathrm{T}} B$$

是三对角矩阵，即相当于已将$A^{\mathrm{T}}A$三对角化。

分解式（5.1）可用Householder变换实现。将A分块为

$$A = [\underset{1}{v_1}, \underset{n-1}{A_1}].$$

第一步，先计算m阶Householder变换P_1使得

$$P_1 v_1 = \delta_1 e_1 \quad (\delta_1 \in \mathbb{R}, e_1 \in \mathbb{R}^m),$$

并形成

$$\begin{matrix} 1 \\ m-1 \end{matrix}\begin{bmatrix} u_1^{\mathrm{T}} \\ \tilde{A}_1 \end{bmatrix} = P_1 A_1;$$

再计算 $n-1$ 阶 Householder 变换 $\tilde{H}_1$ 使得

$$\tilde{H}_1 u_1 = \gamma_2 e_1 \quad (\gamma_2 \in \mathbf{R}, e_1 \in \mathbf{R}^{n-1}),$$

并形成

$$[\underset{1}{v_2}, \underset{n-2}{A_2}] = \tilde{A}_1 \tilde{H}_1.$$

然后，对 $k=2,3,\cdots,n-2$ 依次进行：

(a) 计算 $m-k+1$ 阶 Householder 变换 $\tilde{P}_k$ 使得

$$\tilde{P}_k v_k = \delta_k e_1 \quad (\delta_k \in \mathbf{R},\ e_1 \in \mathbf{R}^{m-k+1}),$$

并形成

$$\begin{matrix} 1 \\ m-k \end{matrix}\underset{n-k}{\begin{bmatrix} u_k^{\mathrm{T}} \\ \tilde{A}_k \end{bmatrix}} = \tilde{P}_k A_k;$$

(b) 计算 $n-k$ 阶 Householder 变换 $\tilde{H}_k$ 使得

$$\tilde{H}_k u_k = \gamma_{k+1} e_1 (\gamma_{k+1} \in \mathbf{R}, e_1 \in \mathbf{R}^{n-k}),$$

并形成

$$[\underset{1}{v_{k+1}}, \underset{n-k-1}{A_{k+1}}] = \tilde{A}_k \tilde{H}_k.$$

进行到 $k=n-2$ 之后，再计算 $m-n+2$ 阶 Householder 变换 $\tilde{P}_{n-1}$ 使得

$$\tilde{P}_{n-1} v_{n-1} = \delta_{n-1} e_1,$$

并形成

$$\begin{matrix} 1 \\ m-n+1 \end{matrix}\underset{1}{\begin{bmatrix} \gamma_n \\ v_n \end{bmatrix}} = \tilde{P}_{n-1} A_{n-1};$$

然后再计算 $m-n+1$ 阶 Householder 变换 $\tilde{P}_n$ 使得

$$\tilde{P}_n v_n = \delta_n e_1.$$

现令

$$P_k = \mathrm{diag}(I_{k-1}, \tilde{P}_k), \quad k=2,\cdots,n,$$

$$H_k = \mathrm{diag}(I_k, \tilde{H}_k), \quad k = 1, 2, \cdots, n-2,$$

$$U_1 = P_1 P_2 \cdots P_n, \quad V_1 = H_1 H_2 \cdots H_{n-2},$$

$$B = \begin{bmatrix} \delta_1 & \gamma_2 & & & 0 \\ & \delta_2 & \gamma_3 & & \\ & & \ddots & \ddots & \\ & 0 & & \ddots & \gamma_n \\ & & & & \delta_n \end{bmatrix},$$

则有

$$U_1^{\mathrm{T}} A V_1 = \begin{bmatrix} B \\ 0 \end{bmatrix} \begin{matrix} n \\ m-n \end{matrix},$$

即实现了分解（5.1）.

这里我们略去算法的详细叙述，请读者作为练习将上述约化过程总结为一个实用的二对角化算法.

将 A 二对角化(即相当于将 $A^{\mathrm{T}}A$ 三对角化)之后,下一步的任务就是对三对角阵

$$T = B^{\mathrm{T}} B$$

进行带 Wilkinson 位移的对称 QR 迭代. 当然，这一步我们亦希望不明确地将 T 计算出来就可进行.

大家知道，应用对称 QR 迭代于 $T = B^{\mathrm{T}}B$ 上的第一步是选取位移 μ，即取矩阵 T 之右下角 2×2 主子阵

$$\begin{bmatrix} \delta_{n-1}^2 + \gamma_{n-1}^2 & \delta_{n-1}\gamma_n \\ \delta_{n-1}\gamma_n & \delta_n^2 + \gamma_n^2 \end{bmatrix}$$

靠近 $\gamma_n^2 + \delta_n^2$ 的特征值作为位移 μ. 这一步当然可以不事先将 $T = B^{\mathrm{T}}B$ 计算好就可进行.

迭代的第二步，就是确定 Givens 变换 $G_1 = G(1,2,\theta)$，其中 $c = \cos\theta$ 和 $s = \sin\theta$ 满足

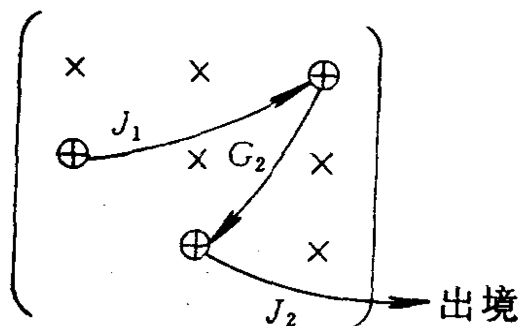
$$\begin{bmatrix} c & -s \\ s & c \end{bmatrix} \begin{bmatrix} \delta_1^2 - \mu \\ \delta_1\gamma_2 \end{bmatrix} = \begin{bmatrix} * \\ 0 \end{bmatrix},$$

这里 $\delta_1^2-\mu$ 和 $\delta_1\gamma_2$ 是 $T-\mu I$ 的第一列位于（1,1）和（1,2）位置的仅有的两个非零元。这一步亦不需事先将 T 明确求出。

迭代的第三步就是确定正交矩阵 Q 使 $Q^{\mathrm{T}}(G_1^{\mathrm{T}}TG_1)Q$ 是对称三对角阵，且 $Qe_1=e_1$。这一步，为了避免 $T=B^{\mathrm{T}}B$ 的计算，只需求正交矩阵 P 和 Q 使 $P^{\mathrm{T}}(BG_1)Q$ 是二对角阵，且 $Qe_1=e_1$ 即可（当然，这需要 T 不可约的条件）。这是容易办到的。可用类似于将 $G_1^{\mathrm{T}}TG_1$ 三对角化的"驱逐出境"法来处理。为了叙述简单起见，下面就 $n=3$ 的情形，说明如何将 BG_1 二对角化。易知此时 BG_1 具有如下形状

$$BG_1=\begin{bmatrix}\times & \times & \\ \oplus & \times & \times \\ & & \times\end{bmatrix},$$

即只有(2,1)位置上出现了一个我们不希望有的非零元素。于是，我们可以左乘一个（1,2）坐标平面的 Givens 变换 J_1 消去这一非零元素；但这样做又会在(1,3）位置上出一个非零元素。因此，我们又需右乘一个（2,3）坐标平面的 Givens 变换 G_2 将（1,3）位置上的非零元素化为零；这又会在（3,2)位置上出现一个非零元素，再左乘一个（2,3）坐标平面的 Givens 变换 J_2 将这一非零元素化为零。这样我们就完成了 $n=3$ 时的 BG_1 的二对角化任务。用图示的方式可将上述约化过程形象地表述如下：



对于一般的 n，用完全类似的方法可确定 $2n-3$ 个 Givens 变换 $J_1,G_2,J_2,G_3,\cdots,G_{n-1},J_{n-1}$ 将 BG_1 中不受欢迎的元素"$\oplus$"驱

逐出境，即使

$$J_{n-1}J_{n-2}\cdots J_1(BG_1)G_2\cdots G_{n-1}$$

是二对角阵，而且这样得到的 $G_2,\cdots,G_{n-1}$满足

$$(G_2\cdots G_{n-1})e_1=e_1.$$

这样，我们就得到了计算奇异值分解的最基本的迭代算法。

算法5.1

(1) 输入二对角矩阵 B 的对角元素 $\delta_1,\cdots,\delta_n$ 和次对角元素 $\gamma_2,\cdots,\gamma_n$。

(2) $d:=[(\delta_{n-1}^2+\gamma_{n-1}^2)-(\gamma_n^2+\delta_n^2)]/2$,

$$\mu:=\frac{(\gamma_n^2+\delta_n^2)-(\delta_{n-1}^2\gamma_n^2)}{d+\text{sign}(d)\sqrt{d^2+\delta_{n-1}^2\gamma_n^2}},$$

$x:=\delta_1^2-\mu$, $y:=\delta_1\gamma_2$, $k:=1$,

$Q:=I$, $P:=I$.

(3) 计算 $c=\cos\theta$, $s=\sin\theta$ 和 σ 使

$$[x,y]\begin{bmatrix} c & s \\ -s & c\end{bmatrix}=[\sigma,0],$$

$$\begin{bmatrix} x & \gamma_{k+1} \\ y & \delta_{k+1}\end{bmatrix}:=\begin{bmatrix} \delta_k & \gamma_{k+1} \\ 0 & \delta_{k+1}\end{bmatrix}\begin{bmatrix} c & s \\ -s & c\end{bmatrix},$$

$$Q:=QG(k,k+1,\theta).$$

(4) 如果 $k>1$，则 $\gamma_k:=\sigma$；否则进行下一步。

(5) 计算 $c=\cos\theta$, $s=\sin\theta$ 和σ 使

$$\begin{bmatrix} c & s \\ -s & c\end{bmatrix}\begin{bmatrix} x \\ y\end{bmatrix}=\begin{bmatrix} \sigma \\ 0\end{bmatrix},$$

$$\delta_k:=\sigma,$$

$$P:=PG(k,k+1,\theta)^{\mathrm{T}}.$$

(6) 如果$k<n-1$，则

$$\begin{bmatrix} x & y \\ \delta_{k+1} & \gamma_{k+2} \end{bmatrix} := \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} \gamma_{k+1} & 0 \\ \delta_{k+1} & \gamma_{k+2} \end{bmatrix},$$

$k := k+1$，转步（3）；否则，

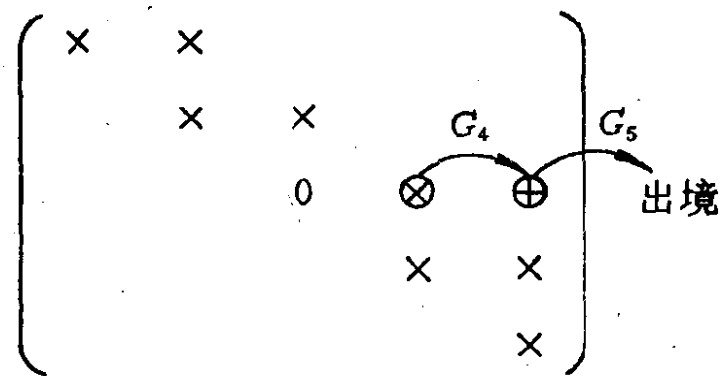
$$\begin{bmatrix} \gamma_n \\ \delta_n \end{bmatrix} := \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} \gamma_n \\ \delta_n \end{bmatrix},$$

迭代结束。

上述算法的导出是在 $T = B^{\mathrm{T}}B$ 不可约的条件下进行的。从 $T = B^{\mathrm{T}}B$ 容易推出，T 不可约的充分必要条件是 δ_i 和 γ_i（除 δ_n 外）都不为零。而当某个 $\gamma_i = 0$ 时，B 具有形状

$$B = \begin{bmatrix} B_1 & 0 \\ 0 & B_2 \end{bmatrix}.$$

因此，可将 B 的奇异值分解问题分解为两个低阶二对角阵的奇异值分解的问题；当某个 $\delta_i = 0$ 时，我们可以给 B 依次左乘 $(i, i+1)$，$(i, i+2)$，…，(i, n) 坐标平面内适当选取的 Givens 变换使 B 变为第 i 行全为零的二对角阵，这一过程可用图示的方式描述如下（例如 $n = 5$，$i = 3$）：



其中 $G_4 = G(3, 4, \theta_4)$，$G_5 = G(3, 5, \theta_5)$。因此，此种情形亦可约化为两个低阶二对角阵的奇异值分解问题。

在实际计算时，当 δ_i 或 γ_j 很小时，就可将 B 分解为两个低

阶二对角阵的奇异值分解问题。通常使用的准则是：当

$$|\delta_i|\leqslant\varepsilon\|B\|_\infty \quad 或 \quad |\gamma_j|\leqslant\varepsilon(|\delta_j|+|\delta_{j-1}|)$$

时，就将δ_i或γ_j视作零，这里ε是一个略大于机器精度的正数。

综述上面的讨论，就可得到现在最流行的计算奇异值分解的算法如下：

算法5.2（SVD 算法）

(1) 输入$A\in\mathbf{R}^{m\times n}(m\geqslant n)$及允许误差$\varepsilon$。

(2) 二对角化：计算 Householder 变换$P_1,\cdots,P_n$，$H_1,\cdots,H_{n-2}$使得

$$(P_1\cdots P_n)^{\mathrm T}A(H_1\cdots H_{n-2})=\begin{bmatrix}B\\0\end{bmatrix}\begin{matrix}n\\m-n\end{matrix},$$

其中

$$B=\begin{bmatrix}\delta_1 & \gamma_2 & & 0\\ & \ddots & \ddots & \\ & 0 & \ddots & \gamma_n\\ & & & \delta_n\end{bmatrix};$$

$$U:=P_1P_2\cdots P_n, \qquad V:=H_1H_2\cdots H_{n-2}.$$

(3) 收敛性检验：

(i) 将所有满足

$$|\gamma_j|\leqslant\varepsilon(|\delta_j|+|\delta_{j-1}|)$$

的γ_j置零；

(ii) 如果$\gamma_j=0$，$j=2,\cdots,n$，则输出有关信息结束；否则，$\gamma_1:=0$，确定正整数$p<q$，使得

$$\gamma_p=\gamma_{q+1}=\cdots=\gamma_n=0, \qquad \gamma_j\neq0,\ p<j\leqslant q;$$

(iii) 如果存在i满足$p\leqslant i\leqslant q-1$使得

$$|\delta_i|\leqslant\varepsilon\|B\|_\infty,$$

则$\delta_i:=0$，$x:=\gamma_{i+1}$，$y:=\delta_{i+1}$，$\gamma_{i+1}:=0$，$l:=1$，转步（iv），否则转步（4）。

(iv) 确定$c=\cos\theta$，$s=\sin\theta$和σ使

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ \sigma \end{bmatrix},$$

$$\delta_{i+l} := \sigma,$$

$$U := UG(i, i+l, \theta)^{\mathrm{T}};$$

（v）如果 $l < q - i$，则

$$x := s\gamma_{i+l+1},\quad \gamma_{i+l+1} := c\gamma_{i+l+1},\quad y := \delta_{i+l+1},$$

$l := l + 1$ 转步（iv），否则转步（i）。

（4）SVD 迭代：应用算法5.1于二对角阵

$$B_1 = \begin{bmatrix} \delta_p & \gamma_{p+1} & & & 0 \\ & \delta_{p+1} & \gamma_{p+2} & & \\ & & \ddots & \ddots & \\ & 0 & & \ddots & \gamma_q \\ & & & & \delta_q \end{bmatrix},$$

得

$$B_1 := P^{\mathrm{T}} B_1 Q,$$

$$U := U\mathrm{diag}(I_p, P, I_{n-p-q}),\quad V := V\mathrm{diag}(I_p, Q, I_{n-p-q}),$$

然后转步（3）。

这一算法可计算任意一个 $m \times n$ 实矩阵 A 的奇异值分解：$A = U\Sigma V^{\mathrm{T}}$。如果用 $\hat{U}, \hat{V}$ 和 $\hat{\Sigma}$ 分别表示 U, V 和 Σ 的计算值，则误差分析的结果表明：

$\hat{U} = W + \Delta U$，　　其中 $W^{\mathrm{T}}W = I_m$，$\|\Delta U\|_2 \leqslant \varepsilon$；

$\hat{V} = Z + \Delta V$，　　其中 $Z^{\mathrm{T}}Z = I_n$，$\|\Delta V\|_2 \leqslant \varepsilon$；

$\hat{\Sigma} = W^{\mathrm{T}}(A + \Delta A)Z$，　　其中 $\|\Delta A\|_2 \leqslant \|A\|_2 \varepsilon$。

这里 ε 一般为机器精度的一个小的倍数。由此可见，这一算法有相当好的数值稳定性；再加上奇异值对扰动的不敏感性，即知利用这一算法可求得相当精确的奇异值。

§6 分而治之法

分而治之法是求实对称三对角矩阵 Schur 分解的一种数值方法，是由Dongarra 和Sorensen 在1987年首先提出的。其基本思想是先将给定的对称三对角阵“分割”成2^k个低阶的对称三对角阵；然后分别求每个低阶的对称三对角阵的 Schur 分解；最后再将这些低阶的Schur 分解“胶合”在一起而得到原矩阵的Schur分解。因此，这一方法特别适用于并行计算。

6.1 分割

设

$$T=\begin{bmatrix} \alpha_1 & \beta_1 & & & 0 \\ \beta_1 & \alpha_2 & \ddots & & \\ & \beta_2 & \ddots & \ddots & \\ & & \ddots & \alpha_{n-1} & \beta_{n-1} \\ 0 & & & \beta_{n-1} & \alpha_n \end{bmatrix} \tag{6.1}$$

是给定的对称三对角矩阵。不妨假定 $n=2m$。定义 $v\in\mathbb{R}^n$ 为

$$v=\begin{bmatrix} e_m^{(m)} \\ \theta e_1^{(m)} \end{bmatrix}, \tag{6.2}$$

其中 θ 为待定的实数。考虑矩阵

$$\tilde{T}=T-\rho vv^{\mathrm{T}},$$

其中 $\rho\in\mathbb{R}$。易知，$\tilde{T}$ 除“中间”的四个元素为

$$\begin{bmatrix} \alpha_m-\rho & \beta_m-\rho\theta \\ \beta_m-\rho\theta & \alpha_{m+1}-\rho\theta^2 \end{bmatrix}$$

外，其余元素与T 完全一样。因此，假如我们取 $\rho\theta=\beta_m$，则

$$T=\begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix}+\rho vv^{\mathrm{T}}, \tag{6.3}$$

其中

$$T_1=\begin{bmatrix}\alpha_1 & \beta_1 & & & 0\\ \beta_1 & \alpha_2 & \ddots & & \\ & \beta_2 & \ddots & \ddots & \\ & & \ddots & \alpha_{m-1} & \beta_{m-1}\\ 0 & & & \beta_{m-1} & \tilde{\alpha}_m\end{bmatrix},$$

$$T_2=\begin{bmatrix}\tilde{\alpha}_{m+1} & \beta_{m+1} & & & 0\\ \beta_{m+1} & \alpha_{m+2} & \ddots & & \\ & \beta_{m+2} & \ddots & \ddots & \\ & & \ddots & \ddots & \beta_{n-1}\\ 0 & & & \beta_{n-1} & \alpha_n\end{bmatrix},$$

这里

$$\tilde{\alpha}_m=\alpha_m-\rho,\qquad \tilde{\alpha}_{m+1}=\alpha_{m+1}-\rho\theta^2.$$

这样我们就把T分割为一个分块矩阵和一个秩 1 矩阵的和。如果需要，还可以对T_1和T_2分别作形如（6.3）的分割。如此下去，就可将T分割为2^k块。

6.2 胶合

假定我们已经求得T_1和T_2的实 Schur 分解

$$Q_1^{\mathrm{T}}T_1Q=D_1 \text{ 和 } Q_2^{\mathrm{T}}T_2Q_2=D_2,$$

其中Q_1和Q_2是m阶正交矩阵，D_1和D_2是对角矩阵。那么，下面的任务就是利用T_1和T_2的 Schur 分解求出T的 Schur 分解，即求正交矩阵V，使

$$V^{\mathrm{T}}TV=\mathrm{diag}(\lambda_1,\cdots,\lambda_n). \tag{6.4}$$

令

$$U=\begin{bmatrix}Q_1 & 0\\ 0 & Q_1\end{bmatrix},$$

则

$$U^{\mathrm{T}}TU = U^{\mathrm{T}}\left(\begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix} + \rho vv^{\mathrm{T}}\right)U$$

$$= D + \rho zz^{\mathrm{T}}, \tag{6.5}$$

其中

$$z = U^{\mathrm{T}}v = \begin{bmatrix} Q_1^{\mathrm{T}}e_m^{(m)} \\ \theta Q_2^{\mathrm{T}}e_1^{(m)} \end{bmatrix}, \qquad D = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}.$$

这样，欲求T的Schur分解问题，就归结为求$D+\rho zz^{\mathrm{T}}$的Schur分解问题。因此下面来考虑如何快速稳定地求矩阵$D+\rho zz^{\mathrm{T}}$的Schur分解。

引理6.1 设$D=\mathrm{diag}(d_1,\cdots,d_n)\in\mathbb{R}^{n\times n}$满足$d_1>d_2>\cdots>d_n$。再假定$\rho\neq0$且$z\in\mathbb{R}^n$的分量均不为零。如果$v\in\mathbb{R}^n$和$\lambda\in\mathbb{R}$满足

$$(D+\rho zz^{\mathrm{T}})v=\lambda v, \qquad v\neq0,$$

则$z^{\mathrm{T}}v\neq0$，且$D-\lambda I$非奇异。

证明 若$z^{\mathrm{T}}v=0$，则$Dv=\lambda v$，$v\neq0$，即λ是D的特征值，v是属于λ的特征向量。而已知D是对角元素互不相同的对角矩阵，故必有某个i使$\lambda=d_i$，而且$v=\alpha e_i$，$\alpha\neq0$。这样便有

$$0=z^{\mathrm{T}}v=\alpha z^{\mathrm{T}}e_i,$$

这与z的分量均不为零的假设矛盾。因此，必有$z^{\mathrm{T}}v\neq0$。

此外，若$D-\lambda I$奇异，则必有某个i，使得$e_i^{\mathrm{T}}(D-\lambda I)=0$；从而有

$$0=e_i^{\mathrm{T}}(D-\lambda I)v=-\rho z^{\mathrm{T}}ve_i^{\mathrm{T}}z,$$

但$\rho z^{\mathrm{T}}v\neq0$，故必有$e_i^{\mathrm{T}}z=0$，这亦与$z$的分量均不为零矛盾。

定理6.1 在引理6.1的假设条件下，再假定$D+\rho zz^{\mathrm{T}}$的Schur分解为

$$V^{\mathrm{T}}(D+\rho zz^{\mathrm{T}})V=\mathrm{diag}(\lambda_1,\cdots,\lambda_n),$$

其中$V=[v_1,\cdots,v_n]$为正交矩阵，$\lambda_1\geq\cdots\geq\lambda_n$，则有

（1）$\lambda_1,\cdots,\lambda_n$正好是函数

$$f(\lambda)=1+\rho z^{\mathrm{T}}(D-\lambda I)^{-1}z$$

的 n 个零点；

(2) 当 $\rho>0$时，有 $\lambda_1>d_1>\lambda_2>d_2>\cdots>\lambda_n>d_n$；

当 $\rho<0$时，有 $d_1>\lambda_1>d_2>\cdots>d_n>\lambda_n$；

(3) 存在常数$\alpha_i\neq 0$，使$v_i=\alpha_i(D-\lambda_i I)^{-1}z, i=1,2,3,\cdots,n$.

证明 由已知条件知，

$$(D+\rho zz^{\mathrm{T}})v_i=\lambda_i v_i,\qquad \|v_i\|_2=1.$$

于是，从引理6.1知，$D-\lambda_i I$ 非奇异；从而有

$$v_i=-\rho z^{\mathrm{T}}v_i(D-\lambda_i I)^{-1}z,\qquad i=1,2,\cdots,n. \tag{6.6}$$

这就证明了 (3)成立，同时也证明了$D+\rho zz^{\mathrm{T}}$的特征值互不相同（否则，如果$\lambda_i=\lambda_j$，则有v_i与v_j线性相关，这与v_i与v_j互相正交矛盾）.

此外，在 (6.6) 两边左乘 z^{T}，并注意到 $z^{\mathrm{T}}v_i\neq 0$，即有

$$1=-\rho z^{\mathrm{T}}(D-\lambda_i I)^{-1}z,$$

即

$$f(\lambda_i)=0,\qquad i=1,2,\cdots,n,$$

这说明 λ_i均是 $f(\lambda)$ 的零点. 下面来证 $f(\lambda)$ 正好有 n 个零点.

设 $z=(\xi_1,\cdots,\xi_n)^{\mathrm{T}}$，则

$$f(\lambda)=1+\rho\left(\frac{\xi_1^2}{d_1-\lambda}+\cdots+\frac{\xi_n^2}{d_n-\lambda}\right).$$

于是有

$$f'(\lambda)=\rho\left(\frac{\xi_1^2}{(d_1-\lambda)^2}+\cdots+\frac{\xi_n^2}{(d_n-\lambda)^2}\right).$$

因此，$f(\lambda)$ 在任意两个相邻的极点 d_i和 d_{i+1}之间是严格单调的：$\rho>0$时，严格增加；$\rho<0$ 时，严格减少. 由此容易知道，$f(\lambda)$ 正好有 n 个零点，而且当$\rho>0$时它们正好分别位于如下的 n 个区间

$$(d_n, d_{n-1}), \cdots, (d_2, d_1), (d_1, \infty);$$

而当$\rho<0$时它们正好分别位于如下的n个区间

$$(-\infty, d_n), (d_n, d_{n-1}), \cdots, (d_2, d_1).$$

由此，立即知定理的(1)和(2)成立。证毕。

从定理6.1可知，在引理6.1的条件下，我们可以按如下两步快速、稳定地求出$D+\rho zz^{\mathrm{T}}$的Schur分解：

第一步　求$f(\lambda)$的零点$\lambda_1, \cdots, \lambda_n$。由每个区间$(d_{i+1}, d_i)$内有且仅有$f(\lambda)$的唯一零点，而且$f(\lambda)$在此区间内严格单调，因此这一步可以应用Newton类型的算法快速、稳定地实现。

第二步　计算

$$v_i = \frac{(D-\lambda_i I)^{-1} z}{\|(D-\lambda_i I)^{-1} z\|_2}, \quad i=1,2,\cdots,n.$$

这一步当然亦可快速、稳定地实现。

对于一般的$D+\rho zz^{\mathrm{T}}$的Schur分解亦可归结为定理6.1所述的情形。为此，我们来构造性地证明下面的结果。

定理6.2　设$D=\mathrm{diag}(d_1, \cdots, d_n) \in \mathbb{R}^{n\times n}$，$z \in \mathbb{R}^n$。则存在正交矩阵$V$和排列$\pi \in \mathscr{S}_n$使得

(1) $V^{\mathrm{T}} z = (\xi_1, \cdots, \xi_r, \underbrace{0, \cdots, 0}_{n-r})^{\mathrm{T}}$满足

$$\xi_i \neq 0, \quad i=1,2,\cdots,r;$$

(2) $V^{\mathrm{T}} D V = \mathrm{diag}(d_{\pi(1)}, \cdots, d_{\pi(n)})$满足

$$d_{\pi(1)} > d_{\pi(2)} > \cdots > d_{\pi(r)}.$$

证明　如果有某两个指标$i<j$使得$d_i=d_j$，则我们可取(i,j)坐标平面内的Givens变换$G(i,j,\theta)$使$G(i,j,\theta)z$的第j个分量为零，而且易证，此时有$G(i,j,\theta)^{\mathrm{T}} D G(i,j,\theta)=D$。这样进行若干步之后，就可找到一个由若干个Givens变换的乘积构成的正交矩阵V_1使得$V_1^{\mathrm{T}} D V_1 = D$，而且$V_1^{\mathrm{T}} z = (\tilde{\xi}_1, \cdots, \tilde{\xi}_n)^{\mathrm{T}}$满足：若$\tilde{\xi}_i \tilde{\xi}_j \neq 0$，$i \neq j$，则必有$d_i \neq d_j$。

然后，再对$V_1^{\mathrm{T}} z$的分量进行若干次两两对换，可使其所有不

为零的分量都位于它的前面，即可找到一个排列方阵P_1，使

$$P_1V_1^{\mathrm T}z=(\xi_{\pi_1(1)},\cdots,\xi_{\pi_1(n)})^{\mathrm T},$$

$$\xi_{\pi_1(i)}\neq 0,\quad i=1,2,\cdots,r,$$

$$\xi_{\pi_1(i)}=0,\quad i=r+1,\cdots,n,$$

其中 $\pi_1\in\mathscr{S}_n$。再由 P_1的取法知，矩阵

$$P_1^{\mathrm T}V_1^{\mathrm T}DV_1P_1=P_1^{\mathrm T}DP_1=\mathrm{diag}(d_{\pi_1(1)},\cdots,d_{\pi_1(n)})$$

的前 r 个对角元 $d_{\pi_1(1)},\cdots,d_{\pi_1(r)}$互不相同。

最后，再对 $d_{\pi_1(1)},\cdots,d_{\pi_1(r)}$进行若干次对换，使它们按从大到小的次序排列，即可找到一个 r 阶排列方阵 $\tilde P_2$，使

$$\tilde P_2^{\mathrm T}\mathrm{diag}(d_{\pi_1(1)},\cdots,d_{\pi_2(r)})\tilde P_2=\mathrm{diag}(\mu_1,\cdots,\mu_r),$$

其中 $\mu_1,\mu_2,\cdots,\mu_r$是由 $d_{\pi_1(1)},\cdots,d_{\pi_2(r)}$ 从大到小排列而得到的，即

$$\mu_1>\mu_2>\cdots>\mu_r.$$

现令

$$V=V_1P_1\mathrm{diag}(\tilde P_2,I_{n-r}),$$

$$V^{\mathrm T}z=(\xi_1,\cdots,\xi_n)^{\mathrm T},$$

$$V^{\mathrm T}DV=\mathrm{diag}(d_{\pi(1)},\cdots,d_{\pi(n)})$$

其中 $\pi\in\mathscr{S}_n$由P_1和$\tilde P_2$决定，则有

$$\xi_i\neq 0,\ i=1,2,\cdots,r,\qquad \xi_{r+1}=\cdots=\xi_n=0,$$

$$d_{\pi(1)}>d_{\pi(2)}>\cdots>d_{\pi(r)}.$$

即定理得证。

由定理6.2可知，对任意的 $D=\mathrm{diag}(d_1,\cdots,d_n)\in\mathbb{R}^n$和$z\in\mathbb{R}^n$可构造出一个正交矩阵$V$，使

$$V^{\mathrm T}(D+\rho zz^{\mathrm T})V=\begin{bmatrix}D_1+\rho ww^{\mathrm T} & 0\\ 0 & D_2\end{bmatrix}\begin{matrix}r\\ n-r\end{matrix},\quad \begin{matrix}r & n-r\end{matrix}$$

其中 $D_1=\mathrm{diag}(d_{\pi(1)},\cdots,d_{\pi(r)})$，$d_{\pi(1)}>\cdots>d_{\pi(r)}$；$D_2=\mathrm{diag}(d_{\pi(r+1)},\cdots,d_{\pi(n)})$；$w=(\xi_1,\cdots,\xi_r)^{\mathrm T}$，$\xi_i\neq 0$，$i=1,2,\cdots,r$，因

此，要求 $D+\rho zz^{\mathrm{T}}$的Schur 分解，我们只需求出 $D_1+\rho ww^{\mathrm{T}}$的 Schur 分解即可，而后者已是定理 6.1 所述的情形，可以快速、稳定地求出。

在实际计算时，当然需事先给定一个准则，来判定何时两数视作相等，何时一个数视作零。一般是取

$$\varepsilon_1=(\|D\|_2+|\rho|\,\|z\|_2)\varepsilon$$

来作为误差限，当$|d_i-d_j|<\varepsilon_1$时，就认为d_i和d_j相等；而当$|\xi_i|<\varepsilon_1$时，就认为ξ_i为零。

作为本节的结束，我们来简要地说明一下如何将这一方法应用于并行计算。为了叙述简单起见，假定我们是在有 4 个处理器的并行机上计算一个$4N$阶的对称三对角矩阵T的Schur 分解。整个计算过程可分为四步：

(1) 分割：

$$T=\begin{bmatrix}T_1 & 0\\ 0 & T_2\end{bmatrix}+\rho vv^{\mathrm{T}},\qquad T_i\in\mathbb{R}^{2N\times 2N},\ v\in\mathbb{R}^{4N};$$

$$T_i=\begin{bmatrix}T_{i1} & 0\\ 0 & T_{i2}\end{bmatrix}+\rho_i w_i w_i^{\mathrm{T}},\qquad T_{ij}\in\mathbb{R}^{N\times N},\ w_i\in\mathbb{R}^{2N}.$$

这步仅需计算少数几个数。

(2) 将 $T_{11},T_{12},T_{21},T_{22}$ 分别分配给四个处理器，去求其 Schur 分解（例如可用对称 QR 方法实现）。

(3) 将T_{11}和T_{12}的Schur 分解以及 T_{21}和T_{22}的Schur 分解分别胶合成 T_1和T_2的Schur 分解。这一步，由于胶合过程主要是求形如 $D+\rho zz^{\mathrm{T}}$的矩阵的特征值和特征向量，而这些特征值的计算基本上是相互独立的，因此亦可分配给四个处理器同时进行。

(4) 再将 T_1和T_2的 Schur 分解胶合成T 的 Schur 分解。这步亦可分配给四个处理器同时进行。

从上面的讨论看出，分而治之法并行的效率是很高的。因此，它适用于在并行机上求解大型对称三对角矩阵的全部特征值和特

征向量。

习　题

1．设 $H\in\mathbb{R}^{n\times n}$ 是不可约的上 Hessenberg 阵。证明：存在对角阵 D，使得 $D^{-1}HD$ 的每个次对角元素皆为 1。$\kappa_2(D)$ 是多少？

2．设 $A=\begin{bmatrix} w & x \\ y & z \end{bmatrix}$ 为实矩阵，并且有特征值 $\lambda\pm i\mu$，其中 μ 不为零。给出确定 $c=\cos\theta$ 和 $s=\sin\theta$ 的稳定算法，使得

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix}^{\mathrm{T}}\begin{bmatrix} w & x \\ y & z \end{bmatrix}\begin{bmatrix} c & s \\ -s & c \end{bmatrix}=\begin{bmatrix} \lambda & \beta \\ \alpha & \lambda \end{bmatrix},$$

其中 $\alpha\beta=-\mu^2$。

3．证明：若对 $H=\begin{bmatrix} w & x \\ y & z \end{bmatrix}\in\mathbb{R}^{2\times 2}$ 进行单步位移的 QR 迭代：

$$H-zI=QR\ \text{（QR 分解）},$$
$$\hat{H}=RQ+zI;$$

则

$$|\hat{h}_{21}|\leqslant|y^2x|/[(w-z)^2+y^2]。$$

4．证明：若给定 $H=H_0$，并且由

$$H_k-\mu_kI=U_kR_k,$$
$$H_{k+1}=R_kU_k+\mu_kI,$$

产生 H_{k+1}，$k=0,1,2,\cdots$，其中 μ_k 是常数，U_k 是正交矩阵，R_k 是上三角矩阵，则

$$(U_0U_1\cdots U_j)(R_j\cdots R_0)=(H-\mu_jI)\cdots(H-\mu_0I)。$$

5．设矩阵

$$T=\begin{bmatrix} T_{11} & T_{12} & T_{13} \\ 0 & T_{22} & T_{23} \\ 0 & 0 & T_{33} \end{bmatrix}\in\mathbb{R}^{n\times n},$$

其中 $T_{22}\in\mathbb{R}^{2\times2}$ 有复共轭特征值，且与 T_{11} 和 T_{22} 的特征值分离。试给出一个求对应于 T_{22} 的特征值的不变子空间的计算方法。

6．令 $x,y\in\mathbb{R}^n$，计算 $c=\cos\theta$ 和 $s=\sin\theta$，使得

$$[x,y]\begin{bmatrix} c & s \\ -s & c \end{bmatrix}$$

的列相互正交。

7．设 λ 是实对称三对角矩阵 T 的特征值。证明：若 λ 的代数重数为 k，则 T 的次对角元素至少有 $k-1$ 个是零。

8．证明：若 $A=B+iC$ 是 Hermite 矩阵，其中 B 和 C 是实矩阵，则

$$M=\begin{bmatrix} B & -C \\ C & B \end{bmatrix}$$

为实对称矩阵。A 与 M 的特征值和特征向量之间有什么关系？

9．设 $A\in\mathbb{C}^{n\times n}$ 是 Hermite 矩阵。给出一个算法，计算酉矩阵 U，使得 U^*AU 为实对称三对角矩阵。

10．设 T 是正定的实对称三对角矩阵.是否按(6.3)式分割之后得到的 T_1 和 T_2 仍是正定的？

第八章　求解实对称特征值问题的同伦方法

§1　同伦算法概述

同伦算法从70年代开始，经过二十多年来的研究，现在已经发展成为一种十分重要的计算方法，得到了广泛的应用，出现了不少行之有效的算法，其中求解实对称特征值问题的同伦算法就是典型的一例。

同伦算法可分为两大类：一类是以微分拓扑学中的Sard定理、原像定理、一维流形分类定理和Euler折线法相结合而发展起来的连续同伦算法；另一类是以代数拓扑学中单纯剖分和单纯逼近为基础而发展起来的单纯同伦算法。这里，只对连续同伦算法作一简略的介绍，有关详细情况可参见文献[13]。

考虑计算非线性映射 $f: \mathbb{R}^n \to \mathbb{R}^n$ 的零点问题。这里，为简化讨论起见，我们假设 f 是光滑映射。

连续同伦方法的基本思想是：先构造 $g: \mathbb{R}^n \to \mathbb{R}^n$ 为一个零点明显的人为光滑映射，并将 f 与 g 用一个连续映射（即所谓的同伦）连结起来；然后再从平凡映射 g 的零点出发逐步过渡到目标映射 f 的零点。通常，连结 f 和 g 的最简单而实用的同伦是：

$$H(x,t) = tf(x) + (1-t)g(x),\quad (x,t) \in \mathbb{R}^n \times [0,1]. \qquad (1.1)$$

显然，H 也是光滑的，而且 $H(x,0) = g(x)$，$H(x,1) = f(x)$。

当然，为了实现从 g 的零点过渡到 f 的零点的目标，必需对人为构造的 g 和 H 有一定的要求。目前，这方面的理论框架已经十分清楚。

下面引进一些记号和定义：

对 $t \in [0,1]$，记 $H_t(\cdot) = H(\cdot, t)$；对 $y \in \mathbb{R}^n$，记

$$H^{-1}(y) = \{(x,t) \in \mathbb{R}^n \times [0,1]:\ H(x,t) = y\},$$

$$H_t^{-1}(y) = \{x \in \mathbb{R}^n: H(x,t) = y\};$$

以 H' 记 H 关于变元 $x_1, \cdots, x_n, t$ 的 $n \times (n+1)$ 阶 Jacobi 矩阵，$H'_{\hat{x}_j}$ 记 H 关于变元 $x_1, \cdots, x_{j-1}, x_j, \cdots, x_n, x_{n+1} = t$ 的 $n \times n$ Jacobi 矩阵，即

$$H' = \begin{bmatrix} \frac{\partial h_1}{\partial x_1} & \frac{\partial h_1}{\partial x_2} & \cdots & \frac{\partial h_1}{\partial x_{n+1}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial h_n}{\partial x_1} & \frac{\partial h_n}{\partial x_2} & \cdots & \frac{\partial h_n}{\partial x_{n+1}} \end{bmatrix},$$

$$H'_{\hat{x}_j} = \begin{bmatrix} \frac{\partial h_1}{\partial x_1} & \cdots & \frac{\partial h_1}{\partial x_{j-1}} & \frac{\partial h_1}{\partial x_{j+1}} & \cdots & \frac{\partial h_1}{\partial x_{n+1}} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{\partial h_n}{\partial x_1} & \cdots & \frac{\partial h_n}{\partial x_{j-1}} & \frac{\partial h_n}{\partial x_{j+1}} & \cdots & \frac{\partial h_n}{\partial x_{n+1}} \end{bmatrix},$$

其中 $x_{n+1} = t$，$H(x,t) = (h_1(x,t), \cdots, h_n(x,t))^{\mathrm{T}}$。

利用上述记号，易知，$H_0 = g$，$H_1 = f$，并且

$$H^{-1}(y) \cap \mathbb{R}^n \times \{t\} = H_t^{-1}(y), \qquad y \in \mathbb{R}^n;$$

特别地

$$H^{-1}(0) \cap \mathbb{R}^n \times \{t\} = H_t^{-1}(0).$$

我们的目标就是希望从已知的 $H_0^{-1}(0) = g^{-1}(0)$ 出发，然后通过 $H^{-1}(0)$ 到达原问题的解集 $H_1^{-1}(0) = f^{-1}(0)$。

定义1.1 设 H 如(1.1)所定义。如果 $(x,t) \in \mathbb{R}^n \times [0,1]$ 满足 $\mathrm{rank}(H'(x,t)) = n$，即 H' 在点 (x,t) 满秩，且当 $t = 0$ 时，还有 $H'_{\hat{t}}$ 在点 (x,t) 满秩，则称 (x,t) 为 H 的一个**正则点**。如果 $y \in \mathbb{R}^n$ 使得 $H^{-1}(y)$ 中的每个点都是 H 的正则点，则称 y 是 H 的一个**正则值**。

利用微分拓扑中的原像定理和一维流形分类定理可以证明（详见文献[12],[13]）：

定理1.1 设$0\in\mathbb{R}^n$是(1.1)所定义的H的正则值。则H的零点集$H^{-1}(0)$由一些互不相交的光滑道路组成，而且每一条这样的道路或是圆周的不与$\mathbb{R}^n\times\{0,1\}$相交的微分同胚像；或是区间的不与$\mathbb{R}^n\times\{0,1\}$相切的微分同胚像，其闭端点必在$\mathbb{R}^n\times\{0,1\}$上。

由定理1.1知，在0是H的正则值的前提下，$H^{-1}(0)$中所包含的道路的种类如图1.1所示。

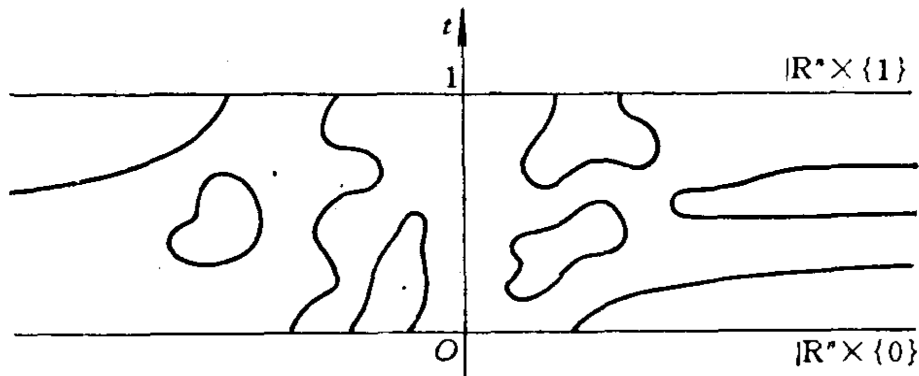


图 1.1

注意，我们的目的是希望从$H_0^{-1}(0)=g^{-1}(0)$的某点出发沿$H^{-1}(0)$的光滑道路到达$H_1=f$的零点。因此，端点不在$\mathbb{R}^n\times\{0\}$上的道路不是我们所关心的；另一方面，端点虽在$\mathbb{R}^n\times\{0\}$上，但却无界的道路，对我们亦无帮助，而且还会带来麻烦。为了消除这种可能性，通常需要H满足如下的**有界性条件**：

存在$\mathbb{R}^n$中的一个有界开集$\mathscr{D}$，使得

$$H^{-1}(0)\cap(\partial\mathscr{D}\times[0,1])=\varnothing,$$

其中$\partial\mathscr{D}$表示$\mathscr{D}$的边界。

在有界性条件成立的前提下，循始点在$\overline{\mathscr{D}}\times[0,1]$($\overline{\mathscr{D}}$表示$\mathscr{D}$的闭包)上的任一条光滑道路可到达$H_1=f$的一个零点，或回到$H_0=g$的另一个零点。在实际计算时，如果只要求算出$f$的一个零点，可取$g$为只有一个零点的光滑映射，此时道路两端都在$\overline{\mathscr{D}}\times\{0\}$上的情形不会出现。

取 s 为道路的弧长参数，则道路上所有 $n+1$ 个变量 $x_1, x_2, \cdots, x_n$，$x_{n+1}=t$ 都可表为参数 s 的函数。设 $0\in \mathbf{R}^n$ 是 H 的正则值，给定初始点 $(x^{(0)}, t^{(0)}) \in H^{-1}(0)$，考虑 Cauchy 问题：

$$\begin{cases} \dfrac{\mathrm{d}x_j}{\mathrm{d}s} = (-1)^{j+1} \det H'_{\hat{x}_j}, \\ x_j(0) = x_j^{(0)}, \end{cases} \quad j=1,2,\cdots,n+1, \qquad (1.2)$$

则我们可以证明：$(x_1(s),\cdots,x_n(s),x_{n+1}(s))$ 是 $H^{-1}(0)$ 中的一条道路的充要条件是它是 Cauchy 问题(1.2)的解(参见文献[12])。

这样，连续同伦方法通过同伦把零点计算问题转化为微分方程的初值问题。因此，我们就可采用 Euler 折线法或预估-校正法求得所需计算的零点。

从上面的讨论，大家已经看到，同伦算法的基本框架已经十分清楚。然而，具体应用时仍有一定的困难。关键是如何选取合适的人为映射 $g(x)$，它既能保证其连接目标映射之零点的光滑道路存在，又能使得数值追踪这样的路径不会付出太大的代价。当然，这并非一件易事，需对所讨论的具体问题作深入透彻的研究，才有可能根据具体问题的特点成功地构造出所需要的映射 $g(x)$。

§2 同伦的构造和性质

考虑如下的对称特征值问题：

$$Ax=\lambda x, \qquad (2.1)$$

这里 A 是已知的 $n\times n$ 实对称矩阵，x 和 λ 是要求的非零实向量和实数。通常称每一对满足(2.1)的向量 x 和实数 λ 为 A 的一对**特征元素**。

由于对任一给定的实对称矩阵，都可以通过标准的三角化过程将其三对角化，故不失一般性可假定 A 是对称三对角矩阵。此外，亦可假定 A 的次对角元素都是非零的，这是因为否则可将原

问题化为几个低阶问题分别考虑．因此，在本章今后的讨论中，如果没有特别说明，我们总假定

$$A=\begin{bmatrix} \alpha_1 & \beta_2 & & & 0 \\ \beta_2 & \alpha_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ 0 & & \ddots & \alpha_{n-1} & \beta_n \\ & & & \beta_n & \alpha_n \end{bmatrix} \text{ 且 } \beta_i \neq 0,\ i=2,\cdots,n.$$

显然，问题(2.1)等价于求非线性映射

$$f(x,\lambda)=\begin{bmatrix} \lambda x - Ax \\ \frac{1}{2}(x^{\mathrm{T}}x-1) \end{bmatrix}, \qquad (x,\lambda)\in \mathbb{R}^n\times R \qquad (2.2)$$

的零点的问题。因此，我们可以应用同伦方法于(2.2)。为此，我们先选取一个实对称三对角矩阵D满足：

(1) $D\in\mathbb{R}^{n\times n}$有$n$个互不相同的特征值，且对应的特征值问题容易求解；

(2) 对任意的$t\in(0,1]$，$A(t)=D+t(A-D)$是次对角元素非零的对称三对角矩阵。

这样的D是容易选取的．例如，我们可取D为对角元素互不相同的对角矩阵．

对于选定的D，定义g：$\mathbb{R}^n\times\mathbb{R}\to\mathbb{R}^n\times\mathbb{R}$为：

$$g(x,\lambda)=\begin{bmatrix} \lambda x - Dx \\ \frac{1}{2}(x^{\mathrm{T}}x-1) \end{bmatrix}, \qquad (x,\lambda)\in \mathbb{R}^n\times \mathbb{R}. \qquad (2.3)$$

然后，再定义H：$\mathbb{R}^n\times\mathbb{R}\times[0,1]\to\mathbb{R}^n\times\mathbb{R}$为：

$$\begin{aligned} H(x,\lambda,t) &= (1-t)g(x,\lambda)+tf(x,\lambda) \\ &= \begin{bmatrix} \lambda x - A(t)x \\ \frac{1}{2}(x^{\mathrm{T}}x-1) \end{bmatrix}, \qquad (x,\lambda,t)\in\mathbb{R}^n\times\mathbb{R}\times[0,1], \end{aligned} \qquad (2.4)$$

其中$A(t)=D+t(A-D)$。

我们将称D为初始矩阵，A为目标矩阵。设D的特征值为$\mu_1,\cdots,\mu_n$，对应的特征向量为$y_1,\cdots,y_n$，则$g(x,\lambda)=0$有且仅有$2n$个解$(\pm y_i,\mu_i)$，$i=1,2,\cdots,n$，这里当然假定$y_i^T y_i=1$。这$2n$个零点是容易求得的。下面我们来讨论这样定义的同伦H的基本性质。

引理2.1 $0\in\mathbb{R}^n\times\mathbb{R}$是(2.4)所定义的$H$的正则值。

证明 容易算出

$$H'=\begin{bmatrix}\lambda I-A(t) & x & (D-A)x\\ x^T & 0 & 0\end{bmatrix}\in\mathbb{R}^{(n+1)\times(n+2)}.$$

现任取$(x,\lambda,t)\in H^{-1}(0)$，则有

$$H(x,\lambda,t)=0,$$

即

$$A(t)x=\lambda x\ \text{且}\ x^T x=1.$$

这也就是说λ是$A(t)$的特征值，x是对应的单位特征向量。由$A(t)$是次对角元素都不为零的实对称三对角矩阵知，$A(t)$的特征值互不相同。因此，必存在正交矩阵Q，使

$$Q^T A(t)Q=\mathrm{diag}(\lambda,\lambda_2,\cdots,\lambda_n),$$

且$Qe_1=x$，$\lambda\neq\lambda_i$，$i=2,\cdots,n$。令

$$U=\mathrm{diag}(\underset{n}{Q},\ \underset{1}{1}).$$

则U是$n+1$阶正交阵，而且

$$U^T\begin{bmatrix}\lambda I-A(t) & x\\ x^T & 0\end{bmatrix}U=\begin{bmatrix}\Lambda & e_1\\ e_1^T & 0\end{bmatrix}$$

是一个非奇异矩阵，其中$\Lambda=\mathrm{diag}(0,\lambda-\lambda_2,\cdots,\lambda-\lambda_n)$。从而，

$$H'_{\hat{t}}=\begin{bmatrix}\lambda I-A(t) & x\\ x^T & 0\end{bmatrix}$$

非奇异。于是H'在点(x,λ,t)满秩，即(x,λ,t)是H的正则点。注意到$(x,\lambda,t)\in H^{-1}(0)$的任意性，即有0是$H$的正则值。

引理2.2 设H由(2.4)所定义，则集合$H^{-1}(0)$是有界的。

证明 任取$(x,\lambda,t)\in H^{-1}(0)$，则有

$$A(t)x=\lambda x \text{ 且 } x^{\mathrm{T}}x=1.$$

于是$\|x\|_2=1$；而且对$t\in[0,1]$，有

$$\begin{aligned}|\lambda| &= \|\lambda x\|_2=\|A(t)x\|_2\\ &\leqslant(1-t)\|D\|_2+t\|A\|_2\\ &\leqslant\|D\|_2+\|A\|_2.\end{aligned}$$

从而$H^{-1}(0)$是有界的。

引理2.3 设H由(2.4)所定义，并假定

$$\Gamma:\ (x(s),\lambda(s),t(s))$$

是$H^{-1}(0)$中的任一条道路，其中s为Γ的弧长参数．则必有

$$\frac{\mathrm{d}t}{\mathrm{d}s}\neq 0,$$

即t是弧长s的单调函数。

证明 由Γ是$H^{-1}(0)$中的道路，故有

$$H(x(s),\lambda(s),t(s))=0,\quad \forall s.$$

对上式两边关于s微分有

$$H'_{\hat{t}}\begin{bmatrix}\dfrac{\mathrm{d}x}{\mathrm{d}s}\\ \\ \dfrac{\mathrm{d}\lambda}{\mathrm{d}s}\end{bmatrix}+\frac{\partial H}{\partial t}\frac{\mathrm{d}t}{\mathrm{d}s}=0. \tag{2.5}$$

从(2.5)易知，若$\dfrac{\mathrm{d}t}{\mathrm{d}s}=0$，则必有$H'_{\hat{t}}$奇异(注意$\left(\dfrac{\mathrm{d}x}{\mathrm{d}s},\dfrac{\mathrm{d}\lambda}{\mathrm{d}s},\dfrac{\mathrm{d}t}{\mathrm{d}s}\right)$ $\neq 0$)。但

$$H'_{\hat{t}}=\begin{bmatrix}\lambda(s)I-A(t(s)) & x(s)\\ x(s)^{\mathrm{T}} & 0\end{bmatrix},$$

而从引理2.1的证明知它应该是非奇异的．这一矛盾说明$\dfrac{\mathrm{d}t}{\mathrm{d}s}=0$

的假定是错误的，从而必有$\frac{\mathrm{d}t}{\mathrm{d}s}\neq 0$，对任意的 s 成立。

根据引理 2.1,2.2 和 2.3 并应用定理 1.1，我们就可得到如下定理。

定理2.1 设 H 是由(2.4)定义的。则 $H^{-1}(0)$由 $2n$ 条互不相交的光滑道路组成，每条道路 Γ 都以D的一对特征元素为起点，而终止于A的一对特征元素，并且可表示为

$$\Gamma:\ (x(t),\lambda(t),t),\qquad t\in[0,1]. \tag{2.6}$$

通常我们称 $H^{-1}(0)$中的每条道路为一条**同伦路径**。对于每条形如(2.6)的同伦路径 Γ，从引理 2.1 的证明知，矩阵

$$B(x(t),\lambda(t))=\begin{bmatrix}\lambda(t)I-A(t) & x(t)\\ x(t)^{\mathrm{T}} & 0\end{bmatrix}$$

对任意的 $t\in[0,1]$都是非奇异的。现沿着 Γ 对 H 微分可得

$$\begin{bmatrix}\frac{\mathrm{d}x}{\mathrm{d}t}\\ \frac{\mathrm{d}\lambda}{\mathrm{d}t}\end{bmatrix}=[B(x(t),\lambda(t))]^{-1}\begin{bmatrix}(A-D)x(t)\\ 0\end{bmatrix},\qquad t\in[0,1]. \tag{2.7}$$

因此，$H^{-1}(0)$中的 $2n$ 条同伦路径，对应于微分方程(2.7)分别以D的 $2n$ 对特征元素$(\pm y_i,\mu_i)(i=1,2,\cdots,n)$为初值的 Cauchy 问题的 $2n$ 条解曲线。于是，我们可以通过求解这 $2n$ 个Cauchy问题（同一个方程的 $2n$ 个不同的初值）而得到所需的 $2n$ 条同伦路径，进而求得A的所有特征值和对应的特征向量。这样，求解常微分方程组初值问题的各种数值方法就可以用来作为计算这$2n$条同伦路径的数值方法。而就我们所关心的问题而言，只需求以$(y_i,\mu_i)(i=1,2,\cdots,n)$为起点的 n 条同伦路径即可，这是由于这 n 条同伦路径的终点已经给出了A的全部特征值和对应的单位特征向量。下一节，我们将根据H的特点详细地讨论如何高效地数值追踪这 n 条同伦路径。

§3 同伦路径的数值追踪

设

$$\Gamma: (x(t), \lambda(t)), \quad t \in [0,1] \tag{3.1}$$

是由(2.4)所定义的同伦 H 所确定的任一条同伦路径，并假定它的起点为D的第 i 对特征元素(y_i, μ_i)。这一节，我们来讨论怎样数值追踪这条同伦路径，以达到求出其终点$(x(1), \lambda(1))$(即A的第 i 对特征元素）的目的。

数值追踪一条同伦路径 Γ 的基本思想是，在$[0,1]$上适当地选取一些点

$$0 = t_0 < t_1 < t_2 < \cdots < t_m = 1,$$

然后从$(x(t_0), \lambda(t_0))$出发，第k步是利用前面 $k-1$ 步所得到的关于$(x(t_i), \lambda(t_i))(i = 0, 1, 2, \cdots, k-1)$的信息使用某种数值方法来确定$(x(t_k), \lambda(t_k))$的近似值，一直追赶到目标$(x(t_m), \lambda(t_m))$为止。通常使用的数值方法就是所谓的**预估-校正法**：首先利用前面所得到的信息给出 $(x(t_k), \lambda(t_k))$的一个初步估计$(\hat{x}(t_k), \hat{\lambda}(t_k))$，即所谓的预估；然后使用收敛速度较快的迭代法以$(\hat{x}(t_k), \hat{\lambda}(t_k))$为初值进行迭代，给出$(x(t_k), \lambda(t_k))$的较精确的近似值，即所谓的校正。

而就我们现在将要追踪的同伦路径 Γ 而言，由于对每个 $t \in [0,1]$，$(x(t), \lambda(t))$实质上是不可约实对称三对角矩阵

$$A(t) = D + t(A - D)$$

的一对特征元，因而我们第 k 步的预估和校正应该是，先依据前面几步得到的信息给出 $A(t_k)$的特征值和相应的特征向量的初步估计，然后利用这一初步估计使用收敛速度较快的迭代法给出$A(t_k)$的特征值和相应的特征向量的更精确的近似值。

下面我们分四步来详细讨论如何具体实现上述的基本想法。

为了避免书写和符号上的麻烦，在下面的讨论中我们记 $r=t_{k-1}$，$s=t_k$，$h=t_{k+1}-t_k$，而且不区别计算值和真值。

3.1 预估

1. 特征值的预估

现假定步长 h 已经确定，并且已知 $(x(r),\lambda(r))$ 和 $(x(s),\lambda(s))$。我们希望给出 $\lambda(s+h)$ 的一个初步估计 $\lambda_0(s+h)$。

先考虑 $s=0$ 的情形。由于初始矩阵 D 的特征元素 (y_j,μ_j) $(j=1,2,\cdots,n)$ 很容易求得，因而可以很容易利用下面的公式求出 $\lambda(t)$ 在 $t=0$ 的前 4 阶导数：

$$\lambda^{(1)}(0)=\left.\frac{\mathrm{d}\lambda}{\mathrm{d}t}\right|_{t=0}=y_i^{\mathrm{T}}(A-D)y_i, \tag{3.2a}$$

$$\lambda^{(2)}(0)=\left.\frac{\mathrm{d}^2\lambda}{\mathrm{d}t^2}\right|_{t=0}=2y_i^{\mathrm{T}}(A-D)\dot{x}(0), \tag{3.2b}$$

$$\lambda^{(3)}(0)=\left.\frac{\mathrm{d}^3\lambda}{\mathrm{d}t^3}\right|_{t=0}=3y_i^{\mathrm{T}}(A-D-\lambda^{(1)}(0)I)\ddot{x}(0), \tag{3.2c}$$

$$\begin{aligned}\lambda^{(4)}(0)=\left.\frac{\mathrm{d}^4\lambda}{\mathrm{d}t^3}\right|_{t=0}&=4y_i^{\mathrm{T}}(A-D-\lambda^{(1)}(0)I)\dddot{x}(0)\\&\quad-6\lambda^{(2)}(0)y_i^{\mathrm{T}}\ddot{x}(0),\end{aligned} \tag{3.2d}$$

其中

$$\dot{x}(0)=G_i(A-D)y_i,$$

$$\ddot{x}(0)=2G_i[A-D-\lambda^{(1)}(0)I]\dot{x}(0)-[\dot{x}(0)^{\mathrm{T}}\dot{x}(0)]y_i,$$

$$\begin{aligned}\dddot{x}(0)=&\,3G_i[(A-D-\lambda^{(1)}(0)I)\ddot{x}(0)-\lambda^{(2)}(0)\dot{x}(0)]\\&-3(\dot{x}(0)^{\mathrm{T}}\ddot{x}(0))y_i,\end{aligned}$$

$$G_i=Y_i\Lambda_iY_i^{\mathrm{T}},$$

$$Y_i=[y_1,\cdots,y_{i-1},y_{i+1},\cdots,y_n],$$

$$\Lambda_i=\mathrm{diag}\left(\frac{1}{\mu_i-\mu_1},\cdots,\frac{1}{\mu_i-\mu_{i-1}},\ \frac{1}{\mu_i-\mu_{i+1}},\cdots,\frac{1}{\mu_i-\mu_n}\right).$$

这些公式很容易利用等式

$$A(t)x(t)=\lambda(t)x(t),\quad x(t)^{\mathrm{T}}x(t)=1,\quad t\in[0,1], \tag{3.3}$$

及 $x(t)$和 $\lambda(t)$的可微性导出. 建议读者作为练习自己推导一下.

因此，对于 $s=0$ 的情形，我们可以用 $\lambda(t)$在 $t=0$ 时的三阶 Taylor 展开式给出 $\lambda(h)$的初步估计：

$$\lambda_0(h)=\lambda(0)+\lambda^{(1)}(0)h+\frac{1}{2}\lambda^{(2)}(0)h^2+\frac{1}{6}\lambda^{(3)}(0)h^3. \tag{3.4}$$

对于 $s\in(0,1)$的情形，由于对任意的 $t\in[0,1]$都有

$$\dot{\lambda}(t)=x(t)^{\mathrm{T}}(A-D)x(t),$$

所以从已知信息很容易求得 $\dot{\lambda}(r)$和 $\dot{\lambda}(s)$，从而可以用以 $\lambda(r)$，$\dot{\lambda}(r)$，$\lambda(s)$，$\dot{\lambda}(s)$ 为插值点的三次 Hermite 插值多项式来给出 $\lambda(s+h)$的估计 $\lambda_0(s+h)$：

$$\lambda_0(s+h)=p(s+h), \tag{3.5}$$

其中

$$p(t)=\lambda(r)+\dot{\lambda}(r)(t-r)+\frac{\lambda(s)-\lambda(r)-\dot{\lambda}(r)(s-r)}{(s-r)^2}(t-r)^2$$

$$+\frac{(s-r)[\dot{\lambda}(r)+\dot{\lambda}(s)]-2[\lambda(s)-\lambda(r)]}{(s-r)^3}(t-r)^2(t-s).$$

2. 特征向量的预估

在求出特征值 $\lambda(s+h)$的初步估计 $\lambda_0(s+h)$之后，我们自然想到，用 $\lambda_0(s+h)$作位移利用反幂法给出特征向量 $x(s+h)$的初步估计 $x_0(s+h)$，而且迭代的初始向量自然可取作前一步计算得到的特征向量 $x(s)$，即

$$x_0(s+h)=y/\|y\|_2, \tag{3.6}$$

其中 y 是方程组

$$[\lambda_0(s+h)I-A(s+h)]y=x(s) \tag{3.7}$$

的解。

3．步长的预估

现在来考虑如何确定步长 h 使得

$$|\lambda(s+h)-\lambda_0(s+h)|\leqslant\varepsilon, \tag{3.8}$$

其中 ε 是事先给定的预估误差(通常不应太小，例如 $\varepsilon\geqslant 10^{-2}$)。

因为我们是用 Hermite 插值多项式来对 $\lambda(s+h)$ 进行估计的，所以由 Hermite 插值多项式的余项公式有

$$\lambda(s+h)-\lambda_0(s+h)=\frac{1}{24}\lambda^{(4)}(\tilde{t})h^2[h+(s-r)]^2, \tag{3.9}$$

其中 $\tilde{t}\in[r,s+h]$。

假如我们已有一估计

$$|\lambda^{(4)}(\tilde{t})|\approx M_s,$$

则从(3.9)知欲使(3.8)成立，只需取 h 使

$$[h+(s-r)]^2h^2M_s=24\varepsilon \tag{3.10}$$

即可。

令

$$f(h)=[h+(s-r)]^2h^2,$$

则 f 的图像如图 3.1 所示。由此易知方程

$$f(h)=24\varepsilon/M_s$$

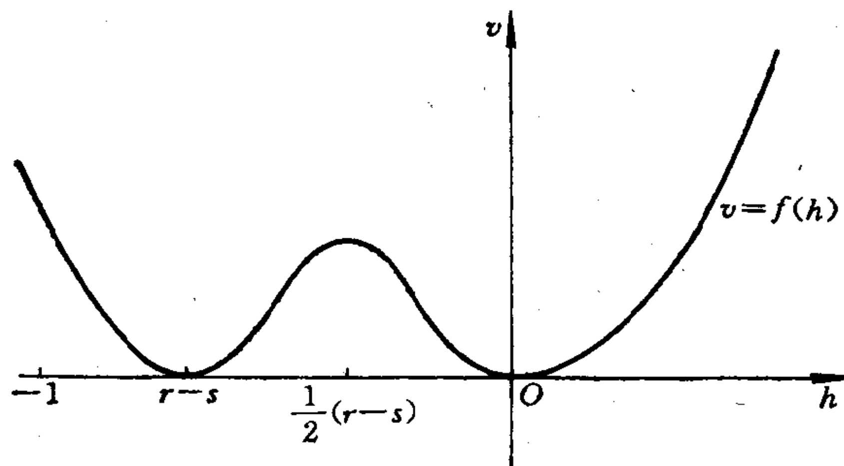


图 3.1

有且仅有唯一的正解。因此，我们可用这一正解 h 来作为步长的预估值。由 f 在 $h>0$ 的严格单调性，这一正解很容易用Newton 迭代法求出，迭代的初值自然可取作前一步的步长 $h_0=s-r$。

现在再来考虑如何给出 M_s 的估计。

对于 $s=0$ 的情形，自然可用 $|\lambda^{(4)}(0)|$ 来作为 M_0 的值，而且初始步长 h 自然应取作

$$h=\left[\frac{24\varepsilon}{M_0}\right]^{1/4}, \tag{3.11}$$

这是由于 $s=0$ 时有

$$\lambda(h)-\lambda_0(h)=\frac{1}{24}\lambda^{(4)}(\bar{t})h^4, \quad \bar{t}\in(0,h)$$

之故。

对于一般的 $s\in(0,1)$，利用前一步对 M_s 的估计值，我们可以通过解方程组(3.10)求出当前的步长 h 的估计，从而可给出 $\lambda(s+h)$ 的估计 $\lambda_0(s+h)$，并通过后面将要介绍的校正法可求得 $\lambda(s+h)$ 的精度很高的值。因此，现在假定 $\lambda_0(s+h)$ 和 $\lambda(s+h)$ 都已求出，我们来给出 M_{s+h} 的估计。

由于

$$\delta=|\lambda(s+h)-\lambda_0(s+h)|=\frac{1}{24}[h+s-r]^2h^2|\lambda^{(4)}(\bar{t})|,$$

故

$$|\lambda^{(4)}(\bar{t})|=24\delta/[[h+s-r]^2h^2].$$

因此，可取

$$M_{s+h}=24\delta/[[h+s-r]^2h^2]. \tag{3.12}$$

3.2 校正

这一步我们需要一个收敛速度较快的求实对称矩阵的一对指定的特征元素的迭代方法。从特征值计算的经典理论可知，Rayleigh 商迭代法是担当此任的最佳候选者。Rayleigh 商迭代法就

是以 Rayleigh 商作位移的反幂法。其基本算法如下：

算法3.1（Rayleigh 商迭代法）

（1）输入实对称矩阵 $B\in\mathbb{R}^{n\times n}$，初始向量 $x_0\in\mathbb{R}^n(\|x_0\|_2=1)$，以及精度要求 ε（通常略大于机器精度）；$k:=1$。

（2）$\nu_k:=x_{k-1}^{\mathrm{T}}Bx_{k-1}$。

（3）求解方程组$(\nu_k I-B)y=x_{k-1}$得 y。

（4）$x_k:=y/\|y\|_2$。

（5）如果 $1/\|y\|_2<\varepsilon$，则输出 x_k 和 ν_k，迭代结束；否则 $k:=k+1$转(2)。

注3.1 （1）在上述迭代中，$\|y\|_2^{-1}<\varepsilon$ 意味着剩余

$$\rho_k=\|(\nu_k I-B)x_k\|_2=1/\|y\|_2<\varepsilon.$$

这蕴含着(x_k,ν_k)已是与 B 非常靠近的一个矩阵的一对特征元素。从而在 B 的特征向量不是十分病态的条件下，(x_k,ν_k)已是 B 的一对很好的近似特征元素。

（2）Rayleigh 商迭代的最终收敛速度是3次的，即若 $\rho_k=\|(\nu_k I-B)x_k\|_2$ 已很小，则

$$\rho_{k+1}=\|(\nu_{k+1}I-B)x_{k+1}\|_2=O(\rho_k^3).$$

基于 Rayleigh 迭代法的快速收敛性，自然应选择其作为校正的数值方法，即在求得 $A(s+h)$的特征向量 $x(s+h)$ 的初步估计 $x_0(s+h)$之后，应对 $B=A(s+h)$和 $x_0=x_0(s+h)$应用算法3.1。

3.3 核查

为了避免在追踪过程中从一条同伦路径跳到另一条同伦路径的现象出现，我们将在追踪过程中不断地检查以保证始终追踪同一条同伦路径。

怎样来判断其是否追踪同一条同伦路径呢？根据不可约对称三对角矩阵的性质知，如果 $\lambda(0)$是 $A(0)$的第 i 个特征值（从小到大排列），那么对任意的 $t\in[0,1]$，$\lambda(t)$亦是 $A(t)$的第 i 个特

征值。因此，我们首先应该利用对称三对角矩阵的 Sturm 序列性质来判断是否追踪同一条同伦路径。

设

$$T=\begin{bmatrix} a_1 & b_2 & & 0 \\ b_2 & a_2 & \ddots & \\ & \ddots & \ddots & b_n \\ 0 & & b_n & a_n \end{bmatrix}$$

为不可约实对称三对角矩阵，并记 $p_i(\lambda)$是T的第 i 阶顺序主子阵的特征多项式。则有下面的三项递推公式：

$$p_0(\lambda)\equiv 1, \qquad p_1(\lambda)=a_1-\lambda,$$

$$p_i(\lambda)=(a_i-\lambda)p_{i-1}(\lambda)-b_i^2 p_{i-2}(\lambda), \quad i=2,3,\cdots,n;$$

而且 p_{i-1} 的零点 $\mu_j^{(i-1)}$ 分隔 p_i 的零点 $\mu_j^{(i)}$：

$$\mu_1^{(i)}<\mu_1^{(i-1)}<\mu_2^{(i)}<\cdots<\mu_{i-1}^{(i-1)}<\mu_i^{(i)}.$$

即 $p_0,p_1,\cdots,p_n$ 构成了一个 sturm 序列。由此易证：

定理3.1 任取 $\mu\in\mathbb{R}^n$，如果记 $G(\mu)$是数列 $p_0(\mu),\cdots,p_n(\mu)$ 的变号数目（如果 $p_i(\mu)=0$，规定 $p_i(\mu)$与 $p_{i-1}(\mu)$同号），则 $G(\mu)$正好是 $p_n(\lambda)$之小于 μ 的根的个数（即T之小于 μ的特征值的个数）。

定理 3.1 的证明作为一个练习请读者自己给出。

令

$$q_i(\lambda)=p_i(\lambda)/p_{i-1}(\lambda), \quad i=1,2,\cdots,n,$$

则 $q_i(\lambda)$可按如下公式递推产生

$$q_1=a_1-\lambda,$$

$$q_i(\lambda)=(a_i-\lambda)-b_i^2/q_{i-1}(\lambda), \quad i=2,\cdots,n.$$

由此易知 $G(\mu)$正好是数列 $q_1(\mu),\cdots,q_n(\mu)$中负数的个数。这样就得到了计算 $G(\mu)$的如下算法。

算法3.2

(1) 输入T的对角元素和次对角元素

$$a_1,\cdots,a_n,b_2,\cdots,b_n,$$

以及实数μ。

(2) $q_1:=a_1-\mu$，$i:=2$。

(3) 如果$q_1<0$，则$G(\mu):=1$；否则$G(\mu):=0$。

(4) 如果$q_{i-1}=0$，则$q_{i-1}:=|b_i|\varepsilon$（$\varepsilon$为机器精度）。

(5) $q_i:=a_i-\mu-b_i^2/q_{i-1}$。

(6) 如果$q_i<0$，则$G(\mu):=G(\mu)+1$。

(7) 如果$i<n$，则$i:=i+1$，转步(4)；否则输出$G(\mu)$，结束。

注3.2 当$q_{i-1}(\mu)=0$时，我们以$|b_i|\varepsilon$来代替了$q_{i-1}(\mu)$，这相当于用$a_{i-1}+|b_i|\varepsilon$代替了a_{i-1}。因此这一算法是数值稳定的。

现在假定对$B=A(s+h)$和$x_0=x_0(s+h)$已用算法3.1求得x_k和ν_k。我们希望判定(x_k,ν_k)是否仍然与同伦路径Γ最接近，亦即希望判定ν_k是否与$A(s+h)$的第i个特征值$\lambda(s+h)$最近。

对$T=A(s+h)$和$\mu=\nu_k$应用算法3.2，可求出$G(\nu_k)$。根据定理3.1，如果ν_k与$\lambda(s+h)$最接近，则$G(\nu_k)$只能等于i或$i-1$。因此，若$G(\nu_k)$不等于i或$i-1$，则表明当前的结果已偏离了所要追踪的路径Γ。但是，当$G(\nu_k)=i$时，我们也只能断定ν_k在$A(s+h)$的第i个特征值$\lambda(s+h)$与第$i+1$个特征值之间，而并不能断定其与$\lambda(s+h)$最近。对于$G(\nu_k)=i-1$亦有同样的问题。因此，我们需要下面的结果。

定理3.2 如果(x_k,ν_k)充分靠近$(x(s+h),\lambda(s+h))$，则有：

(1) 若$x_{k-1}^{\mathrm{T}}x_k>0$，则$\nu_k>\lambda(s+h)$；

(2) 若$x_{k-1}^{\mathrm{T}}x_k<0$，则$\nu_k<\lambda(s+h)$，

其中x_{k-1}为算法3.1中产生x_k的前一个向量。

证明 设$A(s+h)$的特征值和特征向量分别为$\lambda_1,\cdots,\lambda_n$和$u_1,\cdots,u_n$，即

$$A(s+h)u_j=\lambda_j u_j,\quad j=1,2,\cdots,n.$$

且假定$\|u_j\|_2=1$ $(j=1,2,\cdots,n)$，$\lambda_i=\lambda(s+h)$，$u_i=x(s+h)$。

由于 $u_1,\cdots,u_n$ 构成 $\mathbf{R}^n$ 的一组标准正交基，故 x_{k-1} 可按 $u_1,\cdots,u_n$ 展开：

$$x_{k-1}=\sum_{j=1}^{n}\gamma_j u_j,$$

其中

$$\gamma_i=u_j^{\mathrm{T}}x_{k-1},\quad j=1,2,\cdots,n,\qquad \sum_{j=1}^{n}\gamma_j^2=1.$$

由于 x_k 已与 $u_i=x(s+h)$ 很接近，故 γ_i 不会太小。这样便有

$$y=(\nu_k I-A(s+h))^{-1}x_{k-1}=\sum_{i=1}^{n}\frac{\gamma_j}{\nu_k-\lambda_j}u_j.$$

从而有

$$x_{k-1}^{\mathrm{T}}y=\sum_{i=1}^{n}\frac{\gamma_j^2}{\nu_k-\lambda_j}\approx\frac{\gamma_i^2}{\nu_k-\lambda_i}. \tag{3.13}$$

最后的约等于号是由于 ν_k 与 $\lambda_i=\lambda(s+h)$ 已充分靠近而且 γ_i 又不是很小的缘故。再注意到 x_k 与 y 共线，从(3.13)立即得到定理的结论。

将定理3.1和定理(3.2)结合起来，就可给出如下的核查准则：如果下述条件之一成立，

(1) $G(\nu_k)=i$，且 $x_{k-1}^{\mathrm{T}}x_k>0$,

(2) $G(\nu_k)=i-1$，且 $x_{k-1}^{\mathrm{T}}x_k<0$,

则认为现阶段仍在追踪同伦路径 Γ；否则就认为偏离了方向，此时就将当前步长减半并重新开始进行特征值和特征向量的预估和校正。

3.4 同伦算法

综述本节前面的讨论，可得求不可约对称三对角矩阵的第 i 个特征值和对应的单位特征向量的同伦算法如下：

算法3.3（同伦算法）

(1) 输入矩阵 A，D和D的特征元素(y_j,μ_j)，$j=1,2,\cdots,n$，以及预估误差 $\varepsilon>0$（一般 ε 不应太小，例如 $\varepsilon\geqslant10^{-2}$）。

(2) 利用公式(3.2)求出 $\lambda^{(1)}(0),\lambda^{(2)}(0),\lambda^{(3)}(0),\lambda^{(4)}(0)$。

(3) $t:=0$，$\lambda_1:=0$，$d_1:=0$，$\lambda:=\mu_i$，$x:=y_i$，$\lambda_2:=\mu_i$，

$d_2:=\lambda^{(1)}(0)$，$M:=|\lambda^{(4)}(0)|$，$h_0:=0$。

(4) 如果 $t=0$，则

$$h:=[24\varepsilon/M]^{1/4};$$

否则，以 h_0 为初值用 Newton 迭代法求解方程

$$(h+h_0)^2h^2=24\varepsilon/M,$$

得正解 h。

(5) $h:=\min\{h,1-t\}$。

(6) 如果 $t=0$，则

$$\lambda_0:=\mu_i+\lambda^{(1)}(0)h+\frac{1}{2}\lambda^{(2)}(0)h^2+\frac{1}{6}\lambda^{(3)}(0)h^3;$$

否则

$$\lambda_0:=\lambda_1+d_1(h+h_0)+[\lambda_2-\lambda_1+d_1h_0](h+h_0)^2/h_0^2$$
$$+[h_0(d_1+d_2)-2(\lambda_2-\lambda_1)](h+h_0)^2h/h_0^3.$$

(7) 求解线性方程组

$$(\lambda_0I-A(t+h))y=x,$$

得 y，然后

$$x_0:=y/\|y\|_2.$$

(8) 对 $B=A(t+h)$和 x_0 应用算法 3.1，得 x_{k-1},x_k 和 ν_k。

(9) 对 $T=A(t+h)$和 ν_k 应用算法3.2，得 $G(\nu_k)$。

(10) 如果$G(\nu_k)=i$且$x_{k-1}^{\mathrm{T}}x_k>0$，或$G(\nu_k)=i-1$且$x_{k-1}^{\mathrm{T}}x_k<0$，则

$$x:=x_k,\quad \lambda:=\nu_k;$$

否则

$$h:=h/2,$$

转步(6)。

(11) 如果 $t+h=1$，则输出 x,λ，结束；否则

$$\delta:=|\lambda-\lambda_0|,\quad M:=24\delta/(h+h_0)^2h^2,$$

$$h_0:=h,\quad \lambda_1:=\lambda_2,\quad d_1:=d_2,$$

$$\lambda_2:=\lambda,\quad d_2:=x^{\mathrm{T}}(A-D)x,\quad t:=t+h,$$

转步(4)。

这一算法保持原始矩阵 A 始终不变，而且各对特征元素的计算相互独立，因而特别适用于并行计算。此外，李天岩等还进行了数值试验，就他们的例子而言，效果是非常好的，其速度比著名的 QR 方法还要快（平均快 6 到 7 倍），详见文献[48]。但是否真的如此，还有待实践的进一步检验。

习　　题

1．实对称三对角矩阵

$$A=\begin{bmatrix} 2 & 1 & & 0 & \\ 1 & 0 & 3 & & \\ & 3 & 4 & 1 & \\ & & 1 & 1 & 2 \\ & 0 & & 2 & 7 \end{bmatrix}$$

有多少正特征值？

2．Rayleigh 商迭代为：

（1）$\mu_k=x_k^{\mathrm{T}}Ax_k$，

（2）求解方程组$(A-\mu_k I)z_{k+1}=x_k$，得 z_{k+1}，

（3）$x_{k+1}=z_{k+1}/\|z_{k+1}\|_2$，

这里 $k=0,1,2,\cdots$，A 是给定的 n 阶实对称矩阵，$x_0\in\mathbb{R}^n$ 满足 $\|x_0\|_2=1$。证明：

（1）对任意的 k，存在 $\lambda\in\lambda(A)$，使得

$$|\mu_k-\lambda|\leqslant\|z_{k+1}\|_2^{-1};$$

（2）对任意的初始向量 x_0，$\lim\limits_{k\to\infty}\mu_k$ 总是存在的。

3．设 $f(t)=[t+\delta]^2t^2$，其中 $\delta>0$。证明：对于任意的正数 α，方程 $f(t)=\alpha$ 有唯一的正根。并给出求此正根的数值方法。

4．证明定理3.1。

第九章 Lanczos 方法

§1 Lanczos 迭代及其基本性质

这一章，我们来介绍适用于求解大型稀疏对称矩阵特征值问题的 Lanczos 方法。

设 $A\in S\mathbf{R}^{n\times n}$。大家知道，求 A 的特征值和特征向量的大多数方法都要先将 A 三对角化，即求一个正交矩阵 Q，使得

$$T=Q^{\mathrm{T}}AQ \tag{1.1}$$

为对称三对角矩阵.实现三对角化的最常用的方法是 Householder 变换法. 但是，当 A 为大型稀疏矩阵时，这一方法变换过程的中间矩阵很快就失去了稀疏性，需要占用大量的内存，这无疑给处理大型稀疏对称矩阵特征值问题带来了一定的困难。当然，对于某些特殊类型的稀疏矩阵 A，应用 Givens 变换来实现三对角化，在一定程度上可以减少稀疏性的损失.但绝大多数情况下，利用任何一种逐步变换的方式来实现三对角化，都是难以保持约化过程的中间矩阵的稀疏性的。因此，要想充分利用稀疏性尽可能减少内存的占有量，就必须采用直接计算 T 和 Q 的元素的方法来实现三对角化。

设分解式(1.1)已求得. 记

$$Q=[q_1,q_2,\cdots,q_n],$$

$$T=\begin{bmatrix}\alpha_1 & \beta_1 & & & 0\\ \beta_1 & \alpha_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ 0 & & \ddots & \alpha_{n-1} & \beta_{n-1}\\ & & & \beta_{n-1} & \alpha_n\end{bmatrix}.$$

比较

$$AQ = QT$$

两边矩阵的每一列，可得

$$Aq_i = \beta_{i-1}q_{i-1} + \alpha_i q_i + \beta_i q_{i+1}, \quad i = 1,2,\cdots,n, \tag{1.2}$$

这里约定 $\beta_0 q_0 = \beta_n q_{n+1} = 0$. 再根据 q_i 的相互正交性，从(1.2)易得

$$\alpha_i = q_i^{\mathrm{T}} A q_i, \tag{1.3}$$

$$\beta_i = q_{i+1}^{\mathrm{T}} A q_i = \|Aq_i - \beta_{i-1}q_{i-1} - \alpha_i q_i\|_2. \tag{1.4}$$

反过来，对任意给定的$q_1 \in \mathbb{R}^n$，$\|q_1\|_2 = 1$，由(1.2)—(1.4)就可递推地产生 q_i, α_i 和 β_i. 具体写出来就是：

$$\begin{aligned}
&\alpha_1 = q_1^{\mathrm{T}} A q_1, \\
&r_i = Aq_i - \alpha_i q_i - \beta_{i-1} q_{i-1} \ (\beta_0 q_0 = 0), \\
&\beta_i = \|r_i\|_2, \\
&q_{i+1} = r_i / \beta_i \ (\beta_i \neq 0), \\
&\alpha_{i+1} = q_{i+1}^{\mathrm{T}} A q_{i+1}, \\
&i = 1,2,\cdots,n-1.
\end{aligned} \tag{1.5}$$

这就是著名的Lanczos迭代，是由 Lanczos 在1950年首先提出的，其中的 q_i 称作 Lanczos 向量.

现记

$$Q_j = [q_1, \cdots, q_j],$$

$$T_j = \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \alpha_2 & \ddots & & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & & \ddots & \alpha_{j-1} & \beta_{j-1} \\ & & & \beta_{j-1} & \alpha_j \end{bmatrix},$$

则将(1.5)写作矩阵形式即有

$$AQ_j = Q_j T_j + r_j e_j^{\mathrm{T}}, \tag{1.6}$$

通常称 T_j 为 j 阶Lanczos 矩阵.

定理1.1 设q_i，α_i，β_i是由Lanczos迭代(1.5)从初始向量q_1出发而产生的。则有：

(1) 存在正整数m使$\beta_m=0$而$\beta_i\neq 0(1\leqslant i\leqslant m-1)$的充分必要条件是

$$m=\dim(\mathscr{K}(A,q_1,n)),$$

其中$\mathscr{K}(A,q_1,n)=\text{span}\{q_1,Aq_1,\cdots,A^{n-1}q_1\}$；

(2) 对任意的j，$1\leqslant j\leqslant \dim(\mathscr{K}(A,q_1,n))$，有

$$Q_j^{\mathrm{T}}Q_j=I_j \text{ 且 } \mathscr{R}(Q_j)=\mathscr{K}(A,q_1,j), \tag{1.7}$$

其中$\mathscr{K}(A,q_1,j)=\text{span}\{q_1,Aq_1,\cdots,A^{j-1}q_1\}$。

证明 先证对任意的j，$1\leqslant j\leqslant n$，只要

$$\beta_i\neq 0,\quad i=1,2,\cdots,j-1,$$

就有(1.7)成立。

对j应用归纳法。

当$j=1$时，显然有上述结论成立。现假定对$j=k$已证上述结论成立，我们来考虑$j=k+1$的情形。

首先在(1.6)两边左乘Q_k^{T}，并令$j=k$，再应用归纳法假定$Q_k^{\mathrm{T}}Q_k=I_k$，可得

$$Q_k^{\mathrm{T}}AQ_k=T_k+Q_k^{\mathrm{T}}r_ke_k^{\mathrm{T}}. \tag{1.8}$$

其次由(1.5)知

$$\alpha_i=q_i^{\mathrm{T}}Aq_i,\qquad i=1,2,\cdots,k. \tag{1.9}$$

而$Q_k^{\mathrm{T}}Q_k=I_k$即为$q_1,\cdots,q_k$相互正交，故有

$$\begin{aligned}\beta_i&=\beta_iq_{i+1}^{\mathrm{T}}q_{i+1}=q_{i+1}^{\mathrm{T}}r_i\\&=q_{i+1}^{\mathrm{T}}(Aq_i-\alpha_iq_i-\beta_{i-1}q_{i-1})\\&=q_{i+1}^{\mathrm{T}}Aq_i,\qquad i=1,2,\cdots,k-1;\end{aligned} \tag{1.10}$$

$$\begin{gathered}q_l^{\mathrm{T}}Aq_i=q_l^{\mathrm{T}}(\beta_iq_{i+1}+\alpha_iq_i+\beta_{i-1}q_{i-1})=0,\\k\geqslant l\geqslant i+2,\qquad i=1,2,\cdots,k-2.\end{gathered} \tag{1.11}$$

从(1.9)—(1.11)即知

$$Q_k^{\mathrm{T}}AQ_k=T_k. \tag{1.12}$$

将(1.12)代入(1.8)，即得

$$Q_k^{\mathrm{T}} r_k = 0. \tag{1.13}$$

而 $r_k = \beta_k q_{k+1}$ 且 $\beta_k \neq 0$，故有

$$Q_k^{\mathrm{T}} q_{k+1} = 0.$$

这样

$$Q_{k+1}^{\mathrm{T}} Q_{k+1} = \begin{bmatrix} Q_k^{\mathrm{T}} \\ q_{k+1}^{\mathrm{T}} \end{bmatrix} [Q_k, q_{k+1}]$$

$$= \begin{bmatrix} Q_k^{\mathrm{T}} Q_k & Q_k^{\mathrm{T}} q_{k+1} \\ q_{k+1}^{\mathrm{T}} Q_k & q_{k+1}^{\mathrm{T}} q_{k+1} \end{bmatrix} = I_{k+1}.$$

由于

$$\mathscr{R}(Q_k) = \mathscr{K}(q_1, A, k),$$

所以

$$q_{k+1} = \beta_k^{-1}(Aq_k - \alpha_k q_k - \beta_{k-1} q_{k-1}) \in \mathscr{K}(A, q_1, k+1).$$

再注意到 $\dim(\mathscr{K}(A, q_1, k+1) \leqslant k+1$，而 $q_1, \cdots, q_{k+1}$ 相互正交，即有

$$\mathscr{R}(Q_{k+1}) = \mathscr{K}(A, q_1, k+1).$$

现在，剩下的只需证明定理的结论(1)成立即可。

先证必要性。设正整数 m 使 $\beta_m = 0$ 而 $\beta_i \neq 0$，$1 \leqslant i \leqslant m-1$。则由(1.6)知

$$AQ_m = Q_m T_m.$$

这也就是说 $\mathscr{R}(Q_m)$ 是 A 的一个不变子空间。但前面已证

$$\mathscr{R}(Q_m) = \mathscr{K}(A, q_1, m),$$

从而必有

$$\mathscr{K}(A, q_1, n) = \mathscr{K}(A, q_1, m),$$

于是

$$\dim(\mathscr{K}(A, q_1, n)) = m.$$

再证充分性。设 $m = \dim(\mathscr{K}(A, q_1, n))$。则对于任意的 i，$1 \leqslant i \leqslant m-1$，必有 $\beta_i \neq 0$。这是由于，否则由必要性所证知，$\dim(\mathscr{K}(A, q_1, n))$ 将严格小于 m。此外，亦必有 $\beta_m = 0$；否则将又有 $\dim(\mathscr{K}(A, q_1, n))$ 严格大于 m。证毕。

这一定理说明，如果Lanczos迭代过程恒有$\beta_i \neq 0(i=1,2,\cdots,n-1)$，则所产生的$Q_n=[q_1,\cdots,q_n]$和$T_n$将满足

$$Q_n^T A Q_n = T_n \quad 和 \quad Q_n^T Q_n = I_n,$$

即实现了三对角化。而Lanczos迭代过程保持原始矩阵A始终不变，并只涉及到A与向量的乘积，因此可以充分利用A的稀疏性。

如果Lanczos迭代过程出现$\beta_j=0$，则对于求解特征值问题而言是十分有利的。这是因为此时T_j的特征值都是A的特征值。然而实际计算时$|\beta_j|$很小的情形都是罕见的。这样，我们就自然希望知道何时T_j的特征值可作为A的近似特征值。这就需要研究T_j的特征元素与A的特征元素之间的关系。为此，先引进几个概念和记号。

从定理1.1可知，Lanczos矩阵T_j，Lanczos向量矩阵Q_j及A之间有如下关系

$$T_i = Q_j^T A Q_j.$$

因此，又称T_j为A在子空间$\mathscr{R}(Q_j)=\mathscr{K}(A,q_1,j)$上的投影。设$T_j$的特征值为

$$\mu_1 \leqslant \mu_2 \leqslant \cdots \leqslant \mu_j,$$

对应的特征向量分别为

$$y_1,\cdots,y_j.$$

当然，假定

$$y_i^T y_l = \delta_{il}.$$

定义

$$z_i = Q_j y_i, \qquad i=1,2,\cdots,j. \tag{1.14}$$

则称μ_i和z_i为A关于Krylov子空间$\mathscr{K}(A,q_1,j)$的**Ritz值**和**Ritz向量**。

对于Ritz向量，容易验证如下性质：

(1) $z_i^T z_k = \delta_{ik}$，$i,k=1,2,\cdots,j$；

(2) $z_i^T A z_k = \mu_i \delta_{ik}$，$i,k=1,2,\cdots,j$；

(3) $\mathrm{span}\{z_1,\cdots,z_j\} = \mathscr{R}(Q_j)$。

定理1.2 设μ_i和z_i是A关于子空间$\mathscr{K}(A,q_1,j)$的Ritz值和对应的Ritz向量。则

$$\|Az_i-\mu_i z_i\|_2=|\beta_j||\eta_{ij}|, \tag{1.15}$$

其中$\eta_{ij}=y_i^{\mathrm{T}}e_j$，即$\eta_{ij}$是$T_j$对应于$\mu_i$的特征向量$y_i$的最后一个分量。

证明 在等式(1.6)两边右乘y_i，即有

$$AQ_iy_i=Q_jT_jy_i+r_je_j^{\mathrm{T}}y_i,$$

注意到$z_i=Q_jy_i$，$T_jy_i=\mu_iy_i$，上式即为

$$Az_i-\mu_iz_i=r_j\eta_{ij}.$$

上式两边取2范数，并注意到$\|r_j\|_2=\beta_j$，便有(1.15)成立。

推论1.1 对于任意的i，必存在A的一个特征值λ使得

$$|\lambda-\mu_i|\leqslant|\beta_j||\eta_{ij}|, \tag{1.16}$$

其中μ_i,β_j和η_{ij}如定理1.2所述。

证明 令$v=Az_i-\mu_iz_i$，则有

$$(A-vz_i^{\mathrm{T}})z_i=\mu_iz_i,$$

即μ_i和z_i是矩阵$A-vz_i^{\mathrm{T}}$的一对特征元素。由Bauer-Fike定理知，必存在A的一个特征值使得

$$|\lambda-\mu_i|\leqslant\|vz_i^{\mathrm{T}}\|_2\leqslant\|v\|_2,$$

再利用定理1.2即得推论1.1的结论。

推论1.1表明，如果计算到某一步j，T_i的特征值μ_i所对应的特征向量y_i的最后一个分量η_{ij}和β_j的乘积$\eta_{ij}\cdot\beta_j$之绝对值很小的话，那么μ_i就是A的某个特征值λ的很好的近似值。因此，这提醒我们在用Lanczos迭代求A的特征值时，有可能在迭代的中途就可求出A的某些特征值的很好的近似值。当然，这对实际计算大型对称特征值问题时是非常有利的。

§2 Kaniel-Paige-Saad 理论

上一节，我们已经通过简单的分析揭示了在一定条件下

Lanczos矩阵T_j的某些特征值是A的某些特征值的很好的近似值。但是，并没有说明它是A的哪几个特征值的近似值，也没有说明随着j的增加，其逼近程度将如何变化。这一节我们就来介绍这方面的一些较深入的结果——Kaniel-Paige-Saad理论。

S.Kaniel于1966年首先给出了Ritz值的一种收敛性估计，不过证明中有些错误；C.C.Paige于1971年在其著名的博士论文中又重新研究了这一问题，不仅给出了Ritz值的收敛性估计，而且亦给出了Ritz向量的收敛性估计；Y.Saad于1980年在Paige工作的基础上对Ritz值的收敛性给出了比Paige更简单的估计。因此，我们称这些结果为Kaniel-Paige-Saad理论。

先引进一些记号和定义。设T_j是j阶Lanczos矩阵，$\mu_1\leqslant\mu_2\leqslant\cdots\leqslant\mu_j$是$T_j$的特征值，对应的特征向量分别为$y_1,\cdots,y_j$，并满足$y_i^{\mathrm{T}}y_k=\delta_{ik}$，$i,k=1,2,\cdots,j$；再设$A$的特征值为$\lambda_1\leqslant\cdots\leqslant\lambda_n$，对应的特征向量分别为$x_1,\cdots,x_n$，$x_i^{\mathrm{T}}x_k=\delta_{ik}$，$i,k=1,2,\cdots,n$。定义

$$\theta_i=\arccos|x_i^{\mathrm{T}}z_i|,\quad i=1,2,\cdots,j,\tag{2.1}$$

其中$z_i=Q_jy_j$是对应的Ritz向量。适当调整x_i的正负号，我们可以假定

$$x_i^{\mathrm{T}}z_i\geqslant 0,\qquad i=1,2,\cdots,j.$$

由第一章定理4.3知，

$$\mathrm{dist}(\mathscr{R}(x_i),\mathscr{R}(z_i))=\sin\theta_i,\tag{2.2}$$

即$\sin\theta_i$的大小反映了$\mathscr{R}(z_i)$与特征子空间$\mathscr{R}(x_i)$之间的远近程度。因此，我们要了解(μ_i,z_i)与特征元素(λ_i,x_i)之间的远近程度，就需考虑$\sin\theta_i$和$|\lambda_i-\mu_i|$的大小程度。于是，下面我们的中心任务就是给出$\sin\theta_i$和$|\lambda_i-\mu_i|$的上界估计。

由于$T_j=Q_j^{\mathrm{T}}AQ_j$，$Q_j^{\mathrm{T}}Q_j=I_j$，故由第一章的定理6.7知，

$$\lambda_i\leqslant\mu_i\leqslant\lambda_{n+i-j},\qquad i=1,2,\cdots,j,\tag{2.3}$$

因此，有

$$\mu_i-\lambda_i\geqslant 0,\qquad i=1,2,\cdots,j.\tag{2.4}$$

从而只需给出$\mu_i - \lambda_i$的上界估计即可。

定理2.1 设θ_i如(2.1)所定义，则

$$\sin^2\theta_i \leqslant \frac{\mu_i - \lambda_i + \sum_{k=1}^{i-1}(\lambda_{i+1} - \lambda_k)\sin^2\theta_k}{\lambda_{i+1} - \lambda_i} \quad (i = 1, \cdots, j). \tag{2.5}$$

证明 将z_i按$x_1, \cdots, x_n$展开，得

$$z_i = \sum_{k=1}^{n} \alpha_{ik} x_k, \quad \sum_{k=1}^{n} \alpha_{ik}^2 = 1.$$

由θ_i的定义知

$$\cos\theta_i = x_i^{\mathrm{T}} z_i = \alpha_{ii}.$$

所以

$$\sin^2\theta_i = 1 - \alpha_{ii}^2 = \sum_{k=1}^{i-1} \alpha_{ik}^2 + \sum_{k=i+1}^{n} \alpha_{ik}^2. \tag{2.6}$$

此外，从Ritz向量的定义易证

$$\mu_i = z_i^{\mathrm{T}} A z_i = \sum_{k=1}^{n} \alpha_{ik}^2 \lambda_k.$$

从而有

$$\mu_i - \lambda_i = \sum_{k=1}^{n} (\lambda_k - \lambda_i) \alpha_{ik}^2,$$

于是

$$\mu_i - \lambda_i + \sum_{k=1}^{i-1} (\lambda_i - \lambda_k) \alpha_{ik}^2 = \sum_{k=i+1}^{n} (\lambda_k - \lambda_i) \alpha_{ik}^2$$

$$\geqslant (\lambda_{i+1} - \lambda_i) \sum_{k=i+1}^{n} \alpha_{ik}^2. \tag{2.7}$$

由(2.6)和(2.7)得

$$\sin^2\theta_i \leqslant \sum_{k=1}^{i-1} \alpha_{ik}^2 + \frac{\mu_i - \lambda_i + \sum_{k=1}^{i-1} \alpha_{ik}^2 (\lambda_i - \lambda_k)}{\lambda_{i+1} - \lambda_i}$$

$$= \frac{\mu_i - \lambda_i + \sum_{k=1}^{i-1} \alpha_{ik}^2 (\lambda_{i+1} - \lambda_k)}{\lambda_{i+1} - \lambda_i}. \tag{2.8}$$

另外，注意到 $\cos\theta_k = z_k^{\mathrm{T}} x_k$，故 x_k 可表示为

$$x_k = z_k \cos\theta_k + u_k \sin\theta_k, \quad k = 1, 2, \cdots, i-1,$$

其中 u_k 是在 x_k 与 z_k 所生成的平面内与 z_k 垂直的单位向量。这样，当 $i \neq k$ 时，有

$$\alpha_{ik} = z_i^{\mathrm{T}} x_k = z_i^{\mathrm{T}} u_k \sin\theta_k,$$

从而有

$$\alpha_{ik}^2 \leqslant \sin^2\theta_k, \quad k = 1, 2, \cdots, k-1. \tag{2.9}$$

将(2.9)代入(2.8)即知(2.5)成立。证毕。

对于任意的 $B \in S\mathbb{R}^{n\times n}$，$v \in \mathbb{R}^n$，$v \neq 0$，定义

$$\rho(v; B) = v^{\mathrm{T}} B v / v^{\mathrm{T}} v, \tag{2.10}$$

即所谓矩阵 B 关于向量 v 的**Rayleigh商**。

引理2.1 对任意的正整数 $k(1 \leqslant k \leqslant j)$，和任意的非零向量 $v \in \mathscr{K}(A, q_1, j) \cap (\mathrm{span}\{z_1, z_{k-1}\})^{\perp}$，有

$$\mu_k - \lambda_k \leqslant \rho(v; A - \lambda_k I). \tag{2.11}$$

证明 由于

$$\rho(v; A - \lambda_k I) = \rho(v, A) - \lambda_k,$$

故只需证 $\rho(v; A) \geqslant \mu_k$ 即可。此外，无妨假定 $\|v\|_2 = 1$。

因

$$v \in \mathscr{K}(A, q_1, j) = \mathrm{span}\{z_1, \cdots, z_j\},$$

且

$$v^{\mathrm{T}} z_i = 0, \quad i = 1, 2, \cdots, k-1,$$

故 v 可表示为

$$v = \sum_{i=k}^{j} \gamma_i z_i, \quad \sum_{i=k}^{j} \gamma_i^2 = 1.$$

于是有

$$\rho(v; A) = v^{\mathrm{T}} A v = \left(\sum_{i=k}^{j} \gamma_i z_i\right)^{\mathrm{T}} A \left(\sum_{i=k}^{j} \gamma_i z_i\right)$$

$$= \sum_{i=k}^{j} \mu_i \gamma_i^2 \geqslant \mu_k \sum_{i=k}^{j} \gamma_i^2 = \mu_k.$$

引理2.2 对任意的$p(\lambda) \in \mathscr{P}_j$和自然数$i \leqslant j$，$z_i$与$p(A)q_1$正交的充要条件是$\mu_i$是$p(\lambda)$的零点，其中$\mathscr{P}_j$表示次数不超过$j$的实系数多项式的全体。

证明 充分性 设$p(\mu_i)=0$。则

$$p(\lambda) = (\lambda - \mu_i) g(\lambda), \quad g(\lambda) \in \mathscr{P}_{j-1}.$$

而$g(\lambda) \in \mathscr{P}_{j-1}$蕴含着

$$g(A) q_1 \in \mathscr{K}(A, q_1, j) = \mathrm{span}\{q_1, \cdots, q_j\},$$

于是，存在$u \in \mathbb{R}^j$，使得

$$g(A) q_1 = Q_j u.$$

这样，

$$\begin{aligned} z_i^{\mathrm{T}} p(A) q_1 &= y_i^{\mathrm{T}} Q_j^{\mathrm{T}} (A - \mu_i I) Q_j u \\ &= y_i^{\mathrm{T}} (T_j - \mu_i I) u = 0. \end{aligned}$$

必要性 设$z_i^{\mathrm{T}} p(A) q_1 = 0$。并将$p(\lambda)$表示为

$$p(\lambda) = (\lambda - \mu_i) h(\lambda) + \alpha, \qquad h(\lambda) \in \mathscr{P}_{j-1}, \ \alpha \in \mathbb{R}.$$

则同样有$v \in \mathbb{R}^j$，使得$h(A) q_1 = Q_j v$。这样，便有

$$0 = z_i^{\mathrm{T}} p(A) q_1 = \alpha y_i^{\mathrm{T}} e_1.$$

但对于T_j，由于其次对角元素β_i均不为零，故易证其任一特征向量的第一个分量均不为零。从而必有$\alpha = 0$，于是μ_i是$p(\lambda)$的零点。

定理2.2 对任意的自然数$k \leqslant j < n$，和

$$q_1 = \sum_{i=1}^{n} \xi_i x_i, \quad \xi_k \neq 0, \quad \sum_{i=1}^{n} \xi_i^2 = 1,$$

有

$$0 \leqslant \mu_k - \lambda_k \leqslant (\lambda_n - \lambda_k) \sum_{i=k+1}^{n} \left(\frac{\xi_i}{\xi_k} \right)^2 \times \prod_{i=1}^{k-1} \left(\frac{\mu_i - \lambda_n}{\mu_i - \lambda_k} \right)^2 \left(\frac{1}{C_{j-k}(1+2r)} \right)^2, \tag{2.12}$$

其中C_{j-k}为$j-k$阶Chebyshev 多项式，$r=\dfrac{\lambda_k-\lambda_{k+1}}{\lambda_{k+1}-\lambda_n}$。

证明 取 $p(\lambda)\in\mathscr{P}_{j-1}$ 为

$$p(\lambda)=\prod_{i=1}^{k-1}(\lambda-\mu_i)g(\lambda),\quad g(\lambda)\in\mathscr{P}_{j-k}(\text{待定}).$$

记 $v=p(A)q_1$。则显然有 $v\in\mathscr{K}(A,q_1,j)$。再根据引理2.2知 $v\in\mathrm{span}\{z_1,\cdots,z_{k-1}\}^{\perp}$。这样，由引理 2.1 即得

$$\mu_k-\lambda_k\leqslant\rho(v;A-\lambda_kI).\tag{2.13}$$

而 $$\rho(v;A-\lambda_kI)=q_1^{\mathrm{T}}p(A)(A-\lambda_kI)p(A)q_1/[q_1^{\mathrm{T}}p(A)^2q_1]$$

$$=\frac{\sum_{i=1}^{n}\xi_i^2p(\lambda_i)^2(\lambda_i-\lambda_k)}{\sum_{i=1}^{n}\xi_i^2p(\lambda_i)^2}$$

$$\leqslant(\lambda_n-\lambda_k)\frac{\sum_{i=k+1}^{n}\xi_i^2p(\lambda_i)^2}{\sum_{i=1}^{n}\xi_i^2p(\lambda_i)^2}$$

$$\leqslant\frac{(\lambda_n-\lambda_k)\sum_{i=k+1}^{n}\xi_i^2p(\lambda_i)^2}{\xi_k^2p(\lambda_k)^2},\tag{2.14}$$

这里假定 $p(\lambda_k)\neq0$。又

$$\sum_{i=k+1}^{n}\xi_i^2p(\lambda_i)^2\leqslant\max_{k+1\leqslant i\leqslant n}p(\lambda_i)^2\sum_{i=k+1}^{n}\xi_i^2$$

$$\leqslant\left[\prod_{l=1}^{k-1}(\lambda_n-\mu_l)^2\sum_{i=k+1}^{n}\xi_i^2\right]\max_{k+1\leqslant i\leqslant n}g(\lambda_i)^2.\tag{2.15}$$

于是将(2.14)和(2.15)代入(2.13)即得

$$\mu_k-\lambda_k\leqslant(\lambda_n-\lambda_k)\left[\sum_{i=k+1}^{n}\frac{\xi_i^2}{\xi_k^2}\right]$$

$$\times\left[\prod_{l=1}^{k-1}\left(\frac{\lambda_n-\mu_l}{\lambda_k-\mu_l}\right)^2\right]\frac{\max\limits_{k+1\leqslant i\leqslant n} g(\lambda_i)^2}{g(\lambda_k)^2}. \qquad (2.16)$$

注意，这里的 $g(\lambda)\in\mathscr{P}_{j-k}$ 是任意的。因此，我们自然希望选择 $g(\lambda)$ 使得

$$\max_{k+1\leqslant i\leqslant n} g(\lambda_i)/g(\lambda_k)$$

尽可能的小。Chebyshev 多项式的良好性质使其正好充当此任。取

$$g(\lambda)=C_{j-k}\left(\frac{2\lambda-\lambda_{k+1}-\lambda_n}{\lambda_n-\lambda_{k+1}}\right),$$

其中C_{j-k}为$j-k$阶 Chebyshev 多项式，则有

$$|g(\lambda_i)|\leqslant 1, \qquad i=k+1,\cdots,n,$$

而

$$g(\lambda_k)^2=[C_{j-k}(1+2r)]^2, \qquad r=\frac{\lambda_{k+1}-\lambda_k}{\lambda_n-\lambda_{k+1}},$$

当$0<r<0.1$时，随着$j-k$的增大而增长的非常之快（当 $0<r<0.1$ 且$(j-k)\sqrt{r}>1$时，$C_{j-k}(1+2r)\approx\frac{1}{2}\exp(2(j-k)\sqrt{r})$）。对这样选取的$g(x)$，由(2.16)立即得到(2.12)。

注2.1 不等式(2.12)的右端可分为三部分：第一部分为 $\sum\limits_{i=k+1}^{n}\xi_i^2\Big/\xi_k^2$，是由初始向量$q_1$决定的，反映了$q_1$在子空间 $\mathrm{span}\{x_{k+1},\cdots,x_n\}$和 $\mathrm{span}\{x_k\}$ 上的正交投影的大小之比；第二部分为$(\lambda_n-\lambda_k)\prod\limits_{i=1}^{k-1}[(\mu_i-\lambda_n)^2/(\mu_i-\lambda_k)^2]$，主要由$\lambda_k$与$T_j$的特征值的分离程度决定；第三部分为$[C_{j-k}(1+2r)]^{-2}$，是关键的一部分，它反映了收敛的快慢程度，当$k$较小且$0<r<0.1$时，随着$j$的增加而减少的非常之快。

因此，定理2.2实质上表明，随着Lanczos 迭代次数的增加，T_j的前几个特征值$\mu_1,\cdots,\mu_k$（k较小）将非常快的收敛到A的前几

个对应的特征值$\lambda_1,\cdots,\lambda_k$；特别，对$k=1$，(2.12)即为

$$0\leqslant\mu_1-\lambda_1\leqslant(\lambda_n-\lambda_1)\mathrm{tg}^2\varphi/[C_{j-1}(1+2r)]^2, \quad (2.17)$$

其中 $\varphi=\arccos|q_1^{\mathrm{T}}x_1|$，由此更容易看出，随着$j$的增加，$\mu_1$与$\lambda_1$之间的差异将迅速减小。

此外，对于$-A$应用定理2.2，可知T_j的几个较大特征值$\mu_{j-k},\cdots,\mu_j$(k较小)将随着j的增加而很快地收敛到A的几个对应的较大特征值$\lambda_{n-k},\cdots,\lambda_n$；特别有

$$0\leqslant\lambda_n-\mu_j\leqslant(\lambda_n-\lambda_1)\mathrm{tg}^2\psi/(C_{j-1}(1+2\sigma))^2, \quad (2.18)$$

其中 $\psi=\arccos|q_1^{\mathrm{T}}x_n|$，$\sigma=(\lambda_n-\lambda_{n-1})/(\lambda_{n-1}-\lambda_1)$，即说明随着$j$的增加，$T_j$的最大特征值$\mu_j$将非常快地收敛到$A$的最大特征值$\lambda_n$。

§3 Lanczos 算法

前两节对Lanczos迭代的理论分析表明，我们可以利用Lanczos迭代来求实对称矩阵A的某些特征值和对应的特征向量，求解过程可归纳为如下四步：

第一步 利用Lanczos迭代（1.5）产生一列对称三对角矩阵T_j，$j=1,2,\cdots,m$；

第二步 对某一$k\leqslant m$，计算T_k的部分特征值或全部特征值；

第三步 选择这些特征值中的某些个作为A的近似特征值；

第四步 如果对应的特征向量亦需要，则对第三步选定的每个特征值μ，求对应的Ritz向量z（即先求$u\neq0$满足$T_ku=\mu u$，再计算$z=Q_ku$)，然后以z作为A对应于μ的近似特征向量。

对于第二步和第四步已有不少快速有效的数值方法可以利用。例如，可用二分法或对称QR方法求T_k的特征值；可用反幂法求T_k对应的特征向量。

关于第一步，利用(1.5)中某些量的等价公式可设计不同的数值方法加以实现。而从数值性态的优劣而论，较好的方法是：

算法3.1

(1) 输入 $A\in S\mathbb{R}^{n\times n}$，$q_1\in\mathbb{R}^n$ ($\|q_1\|_2=1$)。

(2) $u_1:=Aq_1$，$j:=1$。

(3) $\alpha_j:=q_j^{\mathrm{T}}u_j$，$r_j:=u_j-\alpha_jq_j$，$\beta_j:=\|r_j\|_2$。

(4) 如果 $\beta_j=0$，则结束；否则

$q_{j+1}:=r_j/\beta_j$，　$u_{j+1}:=Aq_{j+1}-\beta_jq_j$，

$j:=j+1$，转步 (3)。

这一算法的主要工作量集中在计算矩阵 A 与向量 v 的乘积 Av 上。在实际使用时，应根据 A 的具体特点，设计一个计算 Av 的子程序，使算法3.1的运算量尽可能的少。

如果算法 3.1 是在没有舍入误差的情况下执行，则所得到的 Lanczos 向量 $q_1,\cdots,q_j$ 是相互正交的，而且至多 n 步必然终止。但是，在误差出现的情况下，计算得到的 Lanczos 向量的正交性很快就会失去，有时甚至还是线性相关的。因此，长期以来人们一直认为这一方法是数值不稳定的，很少用于实际计算。直到1971年，C.C.Paige 在他著名的博士论文（见文献[53]）中，通过仔细的舍入误差分析，发现了失去正交性，恰与近似特征值的精度提高有关。之后，人们又做了大量的理论研究和数值试验，充分认识到，Lanczos 方法对于求高阶稀疏对称矩阵的特征值来说，是非常有效的方法。现在，Lanczos 方法已经成为求解大型稀疏对称矩阵特征值问题的最常用的方法之一。

下面我们就来介绍 Paige 的误差分析结果。设算法 3.1 在机器精度为 ε 的计算机上执行了 k 步。记 $\hat{Q}_k=[\hat{q}_1,\cdots,\hat{q}_k]$和

$$\hat{T}_k=\begin{bmatrix}\hat{\alpha}_1 & \hat{\beta}_1 & & & 0\\ \hat{\beta}_1 & \hat{\alpha}_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & 0 & \ddots & \hat{\alpha}_{k-1} & \hat{\beta}_{k-1}\\ & & & \hat{\beta}_{k-1} & \hat{\alpha}_k\end{bmatrix}$$

分别为Lanczos 向量矩阵 $Q_k=[q_1,\cdots,q_k]$和 Lanczos 矩阵 T_k的计

算结果。令

$$\sigma=\|A\|_2,\qquad \alpha\sigma=\|\,|A|\,\|_2, \tag{3.1}$$

$$\varepsilon_0=2(n+4)\boldsymbol{\varepsilon},\qquad \varepsilon_1=2(7+m\alpha)\boldsymbol{\varepsilon}, \tag{3.2}$$

$$\varepsilon_2=\sqrt{2}\max(6\varepsilon_0,\varepsilon_1), \tag{3.3}$$

其中$|A|$表示A的元素取绝对值之后得到的矩阵，m为A的每行所具有的最多的非零元素的个数。

Paige误差分析的主要结果为（见文献[56]）：

定理3.1 如果$k(3\varepsilon_0+\varepsilon_1)\leqslant 1$，$\varepsilon_0\leqslant 1/12$，则有：

(1)
$$A\hat{Q}_k=\hat{Q}_k\hat{T}_k+\hat{\beta}_k\hat{q}_{k+1}e_k^{\mathrm{T}}+F_k, \tag{3.4}$$

其中$F_k=[f_1,\cdots,f_k]$满足

$$\|f_i\|_2\leqslant\sigma\varepsilon_1,\qquad i=1,2,\cdots,k; \tag{3.5}$$

(2)
$$|\hat{q}_i^{\mathrm{T}}\hat{q}_i-1|\leqslant\frac{1}{2}\varepsilon_0, \tag{3.6}$$

$$\hat{\beta}_i\,|\hat{q}_i^{\mathrm{T}}\hat{q}_{i+1}|\leqslant 2\sigma\varepsilon_0,\qquad i=1,2,\cdots,k; \tag{3.7}$$

(3)
$$\hat{T}_kR_k-R_k\hat{T}_k=\hat{\beta}_k\hat{Q}_k^{\mathrm{T}}\hat{q}_{k+1}e_k^{\mathrm{T}}+H_k, \tag{3.8}$$

其中R_k是$\hat{Q}_k^{\mathrm{T}}\hat{Q}_k$的严格的上三角部分，即

$$\hat{Q}_k^{\mathrm{T}}\hat{Q}_k=R_k^{\mathrm{T}}+\mathrm{diag}(\hat{q}_i^{\mathrm{T}}\hat{q}_i)+R_k, \tag{3.9}$$

H_k是上三角矩阵，且满足

$$\|H_k\|_F\leqslant k\sigma\varepsilon_2. \tag{3.10}$$

这一定理的详细证明，读者可参看文献[55]。

由(3.4)可知，重要等式(1.6)已经计算到机器精度；(3.6)说明计算得到的Lanczos向量几乎是单位向量；(3.7)指出在$\hat{\beta}_i$不很小时，$\hat{q}_i$与$\hat{q}_{i+1}$几乎是正交的；(3.8)是对计算得到的Lanczos向量正交性损失程度的定量刻画。

现在我们来考虑误差对于利用Lanczos方法求解特征值问题的影响。大家知道，对于对称三对角矩阵来讲，已有不少相当有效的数值方法使计算得到的特征值达到机器精度。因此，为了避免在下面讨论中符号上的麻烦，我们不妨假定$\hat{T}_k$的特征值和特

征向量已精确地求出，即假定已经得到正交矩阵

$$Y=[y_1,\cdots,y_k]=[\eta_{ij}]\in\mathbb{R}^{k\times k},$$

和对角矩阵

$$\Sigma=\mathrm{diag}(\mu_1,\cdots,\mu_k),\qquad \mu_1<\mu_2<\cdots<\mu_k,$$

使得

$$\hat{T}_kY=Y\Sigma. \tag{3.11}$$

定义

$$z_j=\hat{Q}_ky_j,\qquad j=1,2,\cdots,k. \tag{3.12}$$

用y_j右乘（3.4）的两边，并利用（3.12），可得

$$Az_j-\mu_jz_j=\hat{\beta}_k\hat{q}_{k+1}\eta_{kj}+F_ky_j. \tag{3.13}$$

类似于推论1.1的证明，可得

$$\min_{1\leqslant i\leqslant n}|\lambda_i-\mu_j|\leqslant\frac{\|Az_j-\mu_jz_j\|_2}{\|z_j\|_2}$$

$$\leqslant\frac{|\hat{\beta}_k\eta_{kj}|(1+\varepsilon_0)+\sqrt{k}\,\sigma\varepsilon_1}{\|z_j\|_2}, \tag{3.14}$$

其中λ_i表示A的第 i 个特征值，最后一个不等式用到了(3.13)，(3.5) 和(3.6)。

(3.14)表明，如果$\|z_j\|_2$不是太小，则只要$|\hat{\beta}_k\eta_{kj}|$很小就有μ_j是A的一个很好的近似特征值。

大家知道，在没有误差的情形下，按照（3.12）定义的z_j应该是单位向量．但在我们现在讨论的前提下，由于正交性的损失，确实会出现$\|z_j\|_2$很小的情况．例如，Paige（参见文献[56]）在11位十进制数系的计算机上进行数值试验，观察到了$\|z_j\|_2=10^{-6}$的情形。

这样一来，我们就需弄清：在什么条件下 $|\hat{\beta}_k\eta_{kj}|$会很小；在什么条件下$\|z_j\|_2$又不会太小。

在等式(3.8)两边分别左乘 y_j^{T}和右乘 y_j即可得：

定理3.2 在定理3.1的假设下有

$$z_j^{\mathrm{T}}\hat{q}_{k+1} = -\varepsilon_{jj}/\beta_k\eta_{kj}, \tag{3.15}$$

其中

$$\varepsilon_{jj} = y_j^{\mathrm{T}}H_k y_j. \tag{3.16}$$

注3.1 由(3.10)知

$$|\varepsilon_{jj}| \leqslant k\sigma\varepsilon_2.$$

因此(3.15)说明$|\beta_k\eta_{kj}|$很小的充分必要条件是$\hat{q}_{k+1}$与z_j失去了正交性（即它们差不多是平行的）。

定理3.3 假设条件同定理3.1。则有

$$\left|\|z_i\|_2^2 - 1\right| \leqslant \frac{k^{5/2}\sigma\varepsilon_2}{\min\limits_{i \neq j}|\mu_i - \mu_j|} + \frac{\varepsilon_0}{2}. \tag{3.17}$$

注3.2 (3.17)表明，只要μ_j与其他μ_i不要太靠近，$\|z_j\|_2$就和1很接近。例如，当$\min\limits_{i \neq j}|\mu_i - \mu_j| \geqslant 6k^{5/2}\sigma\varepsilon_2$时，由$\varepsilon_0 < \frac{1}{12}$从(3.17)就可得到

$$0.75 < \|z_j\|_2 < 1.25. \tag{3.18}$$

而条件$\min\limits_{i \neq j}|\mu_i - \mu_j| \geqslant 6k^{5/2}\sigma\varepsilon_2$并不是特别苛刻的，例如，当$\varepsilon_2$的数量级为$10^{-10}$，$\sigma$的数量级为$10^2$，迭代次数$k = 10^2$时，仅需$\mu_j$与$\hat{T}_k$的其他特征值$\mu_i$之间的距离的数量级不小于$10^{-3}$即可。

定理3.3的证明需要用到对称三对角阵的如下一条基本性质。

引理3.1 设$T \in S\mathbb{R}^{k \times k}$是不可约的三对角矩阵，其特征值为$\mu_1 < \mu_2 < \cdots < \mu_k$，对应的单位特征向量作成的正交矩阵为$Y = [y_1, \cdots, y_k] = [\eta_{ij}]$。则对任意的$1 \leqslant i \leqslant k$和$1 \leqslant j \leqslant k$有

$$\eta_{ij}^2 = \prod_{l=1}^{j-1}\frac{\mu_j - \nu_l}{\mu_j - \mu_l} \cdot \prod_{l=j+1}^{k}\frac{\mu_j - \nu_{l-1}}{\mu_j - \mu_l}, \tag{3.19}$$

其中$\nu_1 \leqslant \nu_2 \leqslant \cdots \leqslant \nu_{k-1}$为$T$划去第$i$行和第$i$列所得到的$k-1$阶矩阵的特征值。

证明 设$p(\lambda)$为T的特征多项式，$q(\lambda)$为T划去第i行第i列所得到的$k-1$阶矩阵的特征多项式。

对于任意的$\lambda \notin \lambda(T)$，有

$$(\lambda I-T)^{-1}=\frac{1}{p(\lambda)}\mathrm{adj}(\lambda I-T), \tag{3.20}$$

其中 $\mathrm{adj}(\lambda I-T)$ 表示 $\lambda I-T$ 的伴随矩阵；另一方面，

$$(\lambda I-T)^{-1}=Y\mathrm{diag}\left(\frac{1}{\lambda-\mu_1},\frac{1}{\lambda-\mu_2},\cdots,\frac{1}{\lambda-\mu_k}\right)Y^{\mathrm{T}}. \tag{3.21}$$

从(3.20)和(3.21)可得

$$\mathrm{adj}(\lambda I-T)=Y\mathrm{diag}\left(\frac{p(\lambda)}{\lambda-\mu_1},\cdots,\frac{p(\lambda)}{\lambda-\mu_k}\right)Y^{\mathrm{T}}. \tag{3.22}$$

由于μ_i互不相同，伴随矩阵的元素是λ的连续函数，故在(3.22)两边令λ趋于μ_j，可得

$$\mathrm{adj}(\mu_j I-T)=p'(\mu_j)y_j y_j^{\mathrm{T}}. \tag{3.23}$$

比较(3.23)两边矩阵的第 i 个对角元素，得

$$q(\mu_j)=p'(\mu_j)\eta_{ij}^2. \tag{3.24}$$

由(3.24)立即知(3.19)成立。

推论3.1 假设条件同引理3.1．则对任意的$2\leqslant i\leqslant k$和$1\leqslant j\leqslant k$，以及T之$i-1$阶顺序主子阵的任一特征值$\mu_r^{(i-1)}$，必存在$\theta_{ij}(r)\in[0,1]$和下标$l(r)$，$1\leqslant l(r)\neq j\leqslant k$，使得

$$\eta_{ij}^2=\theta_{ij}(r)\frac{\mu_j-\mu_r^{(i-1)}}{\mu_j-\mu_{l(r)}}. \tag{3.25}$$

证明 由第一章定理6.7知，

$$\mu_1\leqslant\nu_1\leqslant\mu_2\leqslant\cdots\leqslant\nu_{k-1}\leqslant\mu_k.$$

因此，(3.19)右边乘积中的每个因子都大于等于零，而小于等于1．再注意到T的$i-1$阶顺序主子阵的任一特征值都是T划去第i行和第i列所得矩阵的特征值，故必有某个指标$l(r)$使得$\nu_{l(r)}=\mu_r^{(i-1)}(1\leqslant l(r)\leqslant j-1)$ 或 $\nu_{l(r)-1}=\mu_r^{(i-1)}(j+1\leqslant l(r)\leqslant k)$。这样，将(3.19)右边除去因子 $(\mu_j-\mu_r^{(i-1)})/(\mu_j-\mu_{l(r)})$ 之后剩下

的因子的乘积记作 $\theta_{ij}(r)$，则 $\theta_{ij}(r)\in[0,1]$，且有（3.25）成立。

定理3.3的证明 从(3.9)可得

$$\|z_j\|_2^2-1=2y_j^{\mathrm T}R_ky_j+y_j^{\mathrm T}\mathrm{diag}(\hat q_i^{\mathrm T}\hat q_i-1)y_j. \tag{3.26}$$

再由(3.6) 即知（3.26）右边的第二项之绝对值小于 $\frac{\varepsilon_0}{2}$，这样，欲证（3.17）成立，只需证

$$|y_j^{\mathrm T}R_ky_j|\leqslant\frac{k^{5/2}\sigma\varepsilon_2}{2\min\limits_{i\neq j}|\mu_i-\mu_j|} \tag{3.27}$$

即可。

记 $\hat T_t$ 为 $\hat T_k$ 的 t 阶顺序主子阵，$\hat Q_t=[\hat q_1,\cdots,\hat q_t]$ 并假定 $\hat T_t$ 的Schur 分解为

$$\hat T_t=Y_t\Lambda_tY_t^{\mathrm T},$$

其中 $Y_t=[y_1^{(t)},\cdots,y_t^{(t)}]=[\eta_{ij}^{(t)}]\in\mathbb R^{t\times t}$ 正交，$\Lambda_t=\mathrm{diag}(\mu_1^{(t)},\cdots,\mu_t^{(t)})$，$\mu_1^{(t)}<\mu_2^{(t)}<\cdots<\mu_t^{(t)}$。

由 R_k的定义知，它的第 $t+1$列前 t 个元素作成的向量为

$$\hat Q_t^{\mathrm T}\hat q_{t+1}=Y_t(\hat Q_tY_t)^{\mathrm T}\hat q_{t+1}=Y_tb_t,$$

其中 $b_t=(\hat Q_tY_t)^{\mathrm T}\hat q_{t+1}$。由定理3.2知 b_t的第 r 个分量为

$$e_\tau^{\mathrm T}b_t=-\varepsilon_{\tau\tau}^{(t)}/(\hat\beta_t\eta_{t\tau}^{(t)}),$$

其中

$$\varepsilon_{\tau\tau}^{(t)}=y_\tau^{(t)\mathrm T}H_ty_\tau^{(t)}. \tag{3.28}$$

因此，

$$y_j^{\mathrm T}R_ky_j=\sum_{t=1}^{k-1}\eta_{t+1,j}y_j^{\mathrm T}\begin{bmatrix}Y_tb_t\\0\end{bmatrix}$$

$$=-\sum_{t=1}^{k-1}\eta_{t+1,j}\sum_{\tau=1}^{t}\frac{\varepsilon_{\tau\tau}^{(t)}}{\hat\beta_t\eta_{t\tau}^{(t)}}y_j^{\mathrm T}\begin{bmatrix}y_\tau^{(t)}\\0\end{bmatrix}. \tag{3.29}$$

此外，由于

$$\hat T_k\begin{bmatrix}y_\tau^{(t)}\\0\end{bmatrix}=\mu_\tau^{(t)}\begin{bmatrix}y_\tau^{(t)}\\0\end{bmatrix}+\hat\beta_t\eta_{t\tau}^{(t)}e_{t+1}^{(k)}, \tag{3.30}$$

其中 $e_{t+1}^{(k)}=(0,\cdots,0,\underset{t+1}{1},0,\cdots,0)^{\mathrm{T}}\in\mathbb{R}^k$，故有

$$(\mu_i-\mu_\tau^{(t)})y_j^{\mathrm{T}}\begin{bmatrix}y_\tau^{(t)}\\0\end{bmatrix}=\beta_t\eta_{t\tau}^{(t)}\eta_{t+1,j}. \tag{3.31}$$

把(3.31)代入(3.29)即得

$$y_j^{\mathrm{T}}R_ky_j=-\sum_{t=1}^{k-1}\eta_{t+1,j}^2\sum_{\tau=1}^{t}\frac{\varepsilon_{\tau\tau}^{(t)}}{\mu_j-\mu_\tau^{(t)}}. \tag{3.32}$$

再由推论3.1，对每个 t,τ 和 j 存在 $\theta_{t+1,j}(\tau)\in[0,1]$和指标 $t(\tau)\neq j$ 使得

$$\eta_{t+1,j}^2=\theta_{t+1,j}(\tau)\frac{\mu_j-\mu_\tau^{(t)}}{\mu_j-\mu_{t(\tau)}}.$$

代入（3.32）即有

$$y_j^{\mathrm{T}}R_ky_j=-\sum_{t=1}^{k-1}\sum_{\tau=1}^{t}\frac{\varepsilon_{\tau\tau}^{(t)}}{\mu_j-\mu_{t(\tau)}}\theta_{t+1,j}(\tau).$$

所以

$$\begin{aligned}|y_j^{\mathrm{T}}R_ky_j|&\leqslant\sum_{t=1}^{k-1}\sum_{\tau=1}^{t}\frac{|\varepsilon_{\tau\tau}^{(t)}|}{|\mu_j-\mu_{t(\tau)}|}\\&\leqslant\frac{1}{\min\limits_{i\neq j}|\mu_j-\mu_i|}\sum_{t=1}^{k-1}\sum_{\tau=1}^{t}|\varepsilon_{\tau\tau}^{(t)}|.\end{aligned} \tag{3.33}$$

由(3.10)和 Frobenius 范数的酉不变性知，

$$\begin{aligned}\Big(\sum_{\tau=1}^{t}|\varepsilon_{\tau\tau}^{(t)}|\Big)^2&\leqslant t\sum_{\tau=1}^{t}|\varepsilon_{\tau\tau}^{(t)}|^2\leqslant t\|Y_t^{\mathrm{T}}H_tY_t\|_F^2\\&=t\|H_t\|_F^2\leqslant t^3\sigma^2\varepsilon_2^2,\end{aligned}$$

从而

$$\begin{aligned}\Big(\sum_{t=1}^{k-1}\sum_{\tau=1}^{t}|\varepsilon_{\tau\tau}^{(t)}|\Big)^2&\leqslant(k-1)\sum_{t=1}^{k-1}\Big(\sum_{\tau=1}^{t}|\varepsilon_{\tau\tau}^{(t)}|\Big)^2\\&\leqslant(k-1)\sum_{t=1}^{k-1}t^3\sigma^2\varepsilon_2^2\end{aligned}$$

$$= \frac{1}{4}(k-1)^3 k^2 \sigma^2 \varepsilon_2^2 \leqslant \frac{1}{4} k^5 \sigma^2 \varepsilon_2^2. \tag{3.34}$$

把(3.34)代入(3.33)即得(3.27)。亦即(3.17)成立。证毕。

从这一定理证明中的等式（3.30）知，对任意的 $1 \leqslant t < k$ 和 $\hat{T}_t$ 的任意特征值 $\mu_r^{(t)}$，必有 $\hat{T}_k$ 的特征值 μ_j 使得

$$|\mu_j - \mu_r^{(t)}| \leqslant |\hat{\beta}_t \eta_{tr}^{(t)}|. \tag{3.35}$$

这表明，如果 $|\hat{\beta}_t \eta_{tr}^{(t)}|$ 很小，即 $\mu_r^{(t)}$ 已是 A 的很好的近似特征值，则后面继续计算得到的 Lanczos 矩阵 $\hat{T}_k(k>t)$ 必有特征值与 $\mu_r^{(t)}$ 很靠近作为 A 的很好的近似特征值。当然，这对实际计算是有益的，可望 k 愈大，$\hat{T}_k$ 的特征值中就会含有愈多的 A 的较好的近似特征值。实际计算的经验也确实证实了这一点。人们发现：当 k 充分大时，$\hat{T}_k$ 含有 A 的所有相异的特征值。这就是所谓的**Lanczos 现象**。由于误差的影响，其中的 k 可能远远大于矩阵 A 的阶数 n。

到目前为止，Lanczos 现象还没有得到严格的理论证明。在 Cullum 和Willoughby（1985）关于 Lanczos 算法的专著中，利用 Lanczos 迭代与共轭梯度法之间的等价关系，曾对Lanczos 现象作了一些理论上的解释。由于篇幅所限，这里我们不打算对此作详细的介绍，有兴趣的读者可参看文献[29]的第四章。

对于 A 的某一指定的特征值 λ，究竟 k 多大时 $\hat{T}_k$ 才有与其非常靠近的特征值呢？一般来讲，这主要取决于 A 的特征值的分布情况，取决于 λ 与其他特征值的分离程度以及 λ 在 A 的谱区间（即 $[\lambda_{\min}(A), \lambda_{\max}(A)]$）内所处的位置。粗略地讲，位于区间两端且分离较好的特征值 λ，在 $k \ll n$ 时 $\hat{T}_k$ 的特征值内就含有 λ 的很好的近似值；位于区间内部而又与其他特征值分离的不好的特征值 λ，需 $k \gg n$，$\hat{T}_k$ 的特征值中才会有 λ 的较好的近似值；分离较好的特征值较分离较差的特征值较早出现。

因此，当我们只需计算大型稀疏对称矩阵 A 的少数几个两端特征值时，通常只需迭代很少几步（$k \ll n$），$\hat{T}_k$ 的两端特征值就

是A的两端特征值的很好的近似值。根据 Parllet 的经验(参见文献[57]第259页)，当 $n=10^4$时，取 $k=300$，就可求出 10个 A的两端特征值和对应的特征向量的很好的近似值。而当我们需要求A的全部特征值时，一般 k 要远远大于 n 才行。在Cullum和Willoughby（1985）的专著里介绍了大量的数值例子，他们的经验（参见文献[29]第148页)是：对绝大多数矩阵来讲，只需$k\leqslant 3n$，就可求出其几乎全部的不同特征值达到机器精度的近似值。

在实际使用 Lanczos 方法求解对称矩阵的特征值问题时，还有不少实际问题需要解决。而叙述这些问题及其解决方法又需占用大量的篇幅。因此，这里我们不打算作进一步的介绍，有兴趣的读者可参看 Cullum 和 Willoughby 所著专著（文献[29]）的第四章。

§4 求解对称线性方程组的 Lanczos 方法

这一节我们来介绍 Paige 和 Saunders（1975）给出的求解对称线性方程组的Lanczos 方法（参见文献[54]）。

设$A\in S\mathbb{R}^{n\times n}$非奇异，$b\in\mathbb{R}^n$ 非零。我们来考虑线性方程组

$$Ax=b. \tag{4.1}$$

大家知道：在 A 正定的情形，我们可以用共轭梯度法求解(4.1)；如果取初始向量 $x_0=0$，则共轭梯度法第 k 步迭代产生的近似解x_k是二次泛函

$$\varphi(x)=\frac{1}{2}x^{\mathrm{T}}Ax-x^{\mathrm{T}}b$$

在Krylov 子空间 $\mathscr{K}(A,b,k)$ 上的极小点；而且如果假定 $q_1,\cdots,q_k$是$\mathscr{K}(A,b,k)$ 的一组标准正交基，则有

$$x_k=Q_ky_k, \tag{4.2}$$

其中 $Q_k=[q_1,\cdots,q_k]$，y_k是方程组

$$Q_k^T A Q_k y_k = Q_k^T b \tag{4.3}$$

的唯一解。可是，在A非正定的情形，共轭梯度法可能在还没有求出方程组（4.1）的解之前就发生中断。然而，如果我们按照(4.3)和(4.2)来定义方程组(4.1)在子空间$\mathscr{K}(A,b,k)$上的近似解x_k的话，则不管A是否正定，只要方程组(4.3)是相容的，就可进行。此外，当$k=m=\dim(\mathscr{K}(A,b,n))$时，(4.3)必相容，而且还有$Ax_m=b$。

因此，如果我们能够快速有效地选取$\mathscr{K}(A,b,k)$的正交基，并使得方程组(4.3)容易求解的话，就可望得到求解(4.1)的快速有效的方法。根据 Lanczos 迭代所具有的特性可知，我们正好可以利用 Lanczos 迭代来完成这一任务。

现取$q_1=b/\|b\|_2$，并假定已利用 Lanczos 迭代产生了 Lanczos 矩阵

$$T_k=\begin{bmatrix} \alpha_1 & \beta_1 & & & 0 \\ \beta_1 & \alpha_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ 0 & & \ddots & \alpha_{k-1} & \beta_{k-1} \\ & & & \beta_{k-1} & \alpha_k \end{bmatrix}$$

和 Lanczos 向量作成的矩阵$Q_k=[q_1,\cdots,q_k]$。则有$q_1,\cdots,q_k$是$\mathscr{K}(A,b,k)$的一组标准正交基，而且此时由$Q_k^TAQ_k=T_k$和$b=\|b\|_2 q_1$可知方程（4.3）就变成了

$$T_k y_k=\|b\|_2 e_1. \tag{4.4}$$

这是一个系数矩阵为对称三对角阵的方程组，当其相容时，是非常容易求解的。

但是，如果对每个k都要求解（4.4），然后再由（4.2)产生x_k的话，在计算x_{k+1}时就不能充分利用x_k已经得到的信息，这势必带来一定的浪费。能否充分利用x_k的已有信息来计算x_{k+1}呢？这是我们下面将要讨论的中心问题。

现假定(4.4)是相容的。从理论上和数值计算上来看，求解(4.4)的较满意的方法是正交分解法，即先求 T_k 的正交分解

$$T_k = \tilde{L}_k P_k, \tag{4.5}$$

其中 P_k 是正交矩阵，$\tilde{L}_k$ 是下三角矩阵；然后求解方程组

$$\tilde{L}_k \tilde{z}_k = \|b\|_2 e_1, \tag{4.6}$$

其中 $\tilde{z}_k = P_k y_k$。

分解式(4.5)可以通过对 T_k 右乘 $k-1$ 个 Givens 变换来实现。对于 $k=1$，有

$$\tilde{L}_1 = \bar{\gamma}_1, \qquad P_1 = 1,$$

其中 $\bar{\gamma}_1 = \alpha_1$。现假定对某一 k 已由 $k-1$ 个 Givens 变换确定了一个正交矩阵 P_k，使得

$$T_k P_k^{\mathrm{T}} = \begin{bmatrix} \gamma_1 & & & & & \\ \delta_2 & \gamma_2 & & & 0 & \\ \varepsilon_3 & \delta_3 & \ddots & & & \\ & \varepsilon_4 & \ddots & \ddots & & \\ & 0 & \ddots & \delta_{k-1} & \gamma_{k-1} & \\ & & & \varepsilon_k & \delta_k & \bar{\gamma}_k \end{bmatrix} = \tilde{L}_k. \tag{4.7}$$

定义

$$\gamma_k = (\bar{\gamma}_k^2 + \beta_k^2)^{1/2}, \qquad c_k = \frac{\bar{\gamma}_k}{\gamma_k}, \qquad s_k = \frac{\beta_k}{\gamma_k}, \tag{4.8}$$

$$J_k = \begin{bmatrix} I_{k-1} & & \\ & c_k & s_k \\ & s_k & -c_k \end{bmatrix} \in \mathbb{R}^{(k+1)\times(k+1)},$$

则

$$T_{k+1} \begin{bmatrix} P_k^{\mathrm{T}} & 0 \\ 0 & 1 \end{bmatrix} J_k = \begin{bmatrix} L_k & 0 \\ v^{\mathrm{T}} & \bar{\gamma}_{k+1} \end{bmatrix} \begin{matrix} k \\ 1 \end{matrix}, \tag{4.9}$$
$$\qquad\qquad\qquad\qquad\qquad\quad k \qquad 1$$

其中 $v = (0, \cdots, 0, \varepsilon_{k+1}, \delta_{k+1})^{\mathrm{T}} \in \mathbb{R}^k$，$L_k$ 是由 $\tilde{L}_k$ 将 $\bar{\gamma}_k$ 换成 γ_k 之后得到的下三角阵，从而

$$T_{k+1} = \tilde{L}_{k+1} P_{k+1},$$

这里

$$P_{k+1} = J_k^{\mathrm{T}} \begin{bmatrix} P_k & 0 \\ 0 & 1 \end{bmatrix}, \qquad \tilde{L}_{k+1} = \begin{bmatrix} L_k & 0 \\ v^{\mathrm{T}} & \tilde{\gamma}_{k+1} \end{bmatrix}. \quad (4.10)$$

设 $\tilde{z}_k$ 和 z_k 分别是方程组

$$\tilde{L}_k \tilde{z}_k = \|b\|_2 e_1 \quad 和 \quad L_k z_k = \|b\|_2 e_1 \quad (4.11)$$

之解. 则 $\tilde{z}_k$ 和 z_k 的前 $k-1$ 个分量必然是相同的. 现记 $\tilde{z}_k = (\xi_1, \cdots, \xi_{k-1}, \tilde{\xi})^{\mathrm{T}}$ 和 $z_k = (\xi_1, \cdots, \xi_{k-1}, \xi_k)^{\mathrm{T}}$, 则

$$\gamma_k \xi_k = \tilde{\gamma}_k \tilde{\xi}_k, \quad (4.12)$$

且当 $\beta_k \neq 0$ 时, $\gamma_k \neq 0$。再记

$$W_k = Q_k P_k^{\mathrm{T}} = [w_1, w_2, \cdots, w_{k-1}, \tilde{w}_k],$$

则

$$\begin{aligned} W_{k+1} = Q_{k+1} P_{k+1}^{\mathrm{T}} &= [Q_k, q_{k+1}] \begin{bmatrix} P_k^{\mathrm{T}} & 0 \\ 0 & 1 \end{bmatrix} J_k \\ &= [W_k, q_{k+1}] J_k \\ &= [w_1, \cdots, w_{k-1}, w_k, \tilde{w}_{k+1}], \end{aligned}$$

其中

$$\begin{cases} w_k = c_k \tilde{w}_k + s_k q_{k+1}, \\ \tilde{w}_{k+1} = s_k \tilde{w}_k - c_k q_{k+1}. \end{cases} \quad (4.13)$$

定义

$$\tilde{x}_k = [w_1, \cdots, w_k] z_k,$$

则易知

$$\tilde{x}_k = \tilde{x}_{k-1} + \xi_k w_k. \quad (4.14)$$

而从前面的推理可知, 由(4.4)和(4.2)所确定的 x_k 为

$$\begin{aligned} x_k = Q_k y_k &= Q_k P_k^{\mathrm{T}} \tilde{z}_k \\ &= [w_1, \cdots, w_{k-1}, \tilde{w}_k] \tilde{z}_k \\ &= \xi_1 w_1 + \cdots + \xi_{k-1} w_{k-1} + \tilde{\xi}_k \tilde{w}_k \\ &= \tilde{x}_{k-1} + \tilde{\xi}_k \tilde{w}_k. \end{aligned} \quad (4.15)$$

这样一来，从 $P_1^{\mathrm{T}}=1$ 和 $\tilde{w}_1=q_1=b/\|b\|_2$ 出发，就可用(4.13)递推地产生 w_k 和 $\tilde{w}_{k+1}$；再由(4.14)和(4.15)就可逐步产生 x_k。实际计算时，我们可先不去求 x_k，而只用(4.14)递推地产生 $\tilde{x}_k$，直到计算达到一定精度之后，再由(4.15)计算 x_k。因而，这就需要我们寻找如何来判定一个近似解是否达到一定的精度的方法。从理论上讲，当且仅当 Lanczos 迭代产生的 $\beta_k=0$ 时，x_k 才是方程组(4.1)的精确解。而实际计算时，由于误差的出现，即使是 $|\beta_k|$ 很小的情况也是非常罕见的。所以，我们必须寻找其他的判定方法。

上节关于 Lanczos 算法的误差分析的理论表明，等式

$$AQ_k - Q_kT_k = \beta_k q_{k+1}e_k^{\mathrm{T}}$$

可以计算到机器精度；而对上式两边右乘方程组(4.4)的解 y_k 可得

$$Ax_k - b = \beta_k q_{k+1}\eta_k^{(k)},$$

其中 $\eta_k^{(k)}$ 表示 y_k 的最后一个分量；从而

$$\|Ax_k - b\|_2 = |\beta_k\eta_k^{(k)}|. \tag{4.16}$$

因此，我们可用 $|\beta_k\eta_k^{(k)}|$ 的大小来判定 x_k 是否达到精度要求（因在 A 的条件数不是很大时，剩余很小就表明 x_k 已是方程组(4.1)的很好的近似解）。但这里，我们将由(4.15)来确定 x_k，而并不明确地将 y_k 求出来，因而直接用 $|\beta_k\eta_k^{(k)}|$ 的大小判定近似解 x_k 的精度是不实用的。能否不求出 y_k 就可把分量 $\eta_k^{(k)}$ 计算出来呢？答案是肯定的。下面我们就给出不求 y_k 而直接计算 $\eta_k^{(k)}$ 的方法。

由于 T_k 是实对称的，故

$$T_k = T_k^{\mathrm{T}} = P_k^{\mathrm{T}}\tilde{L}_k^{\mathrm{T}},$$

因此当方程(4.4)相容时，有

$$\tilde{L}_k^{\mathrm{T}}y_k = \|b\|_2 P_k e_1.$$

比较上式两边向量的最后一个分量，可得

$$\tilde{\gamma}_k\eta_k^{(k)} = \|b\|_2 s_1 s_2\cdots s_{k-1}.$$

由(4.8)知，只要 $\beta_i \neq 0$，就有 $s_i \neq 0$；从而当 $\beta_i \neq 0$（$i=1,2,\cdots$，

$k-1$)时，则必有 $\gamma_k \neq 0$。因此，有

$$\eta_k^{(k)} = \|b\|_2 s_1 s_2 \cdots s_{k-1} / \gamma_k. \tag{4.17}$$

现在记 $\Delta_k = |\beta_k \eta_k^{(k)}|$，并注意到(4.8)，就有

$$\Delta_k = \|b\|_2 |s_1 \cdots s_{k-1} s_k / c_k|. \tag{4.18}$$

上述推理亦表明，(4.4)相容的充要条件是 $\gamma_k \neq 0$，这也等价于 $\det(T_k) \neq 0$ 或 $c_k \neq 0$。

再注意到：令 $d_0 = 1$，$d_i = \det(T_i)$，则有

$$d_{i+1} = \alpha_{i+1} d_i - \beta_i^2 d_{i-1}, \qquad i = 1, 2, \cdots, k.$$

故当 $\beta_i \neq 0$ $(i = 1, 2, \cdots, k)$时，若 $d_i = 0$，必有 d_{i+1} 和 d_{i-1} 均不为零，亦即 γ_{i+1} 和 γ_{i-1} 均不为零。这样，据(4.18)，当 γ_{k-1} 和 γ_k 都不为零时，有

$$\Delta_k = \Delta_{k-1} \left| \frac{c_{k-1} s_k}{c_k} \right|;$$

而当 $\gamma_k = 0$ 时，

$$\Delta_{k+1} = \Delta_{k-1} \left| \frac{c_{k-1} s_k s_{k+1}}{c_{k+1}} \right|.$$

由此可知，当遇到 $\gamma_k = 0$（即方程组(4.4)不相容）的情形，我们的计算仍可进行下去，直到求得(4.1)的达到精度要求的近似解为止，即可以避免中断的发生。

综述上面的讨论可给出用 Lanczos 方法求解对称线性方程组的算法如下：

算法4.1

(1) 输入 A, b，及精度要求 ε。

(2) $\beta := \|b\|_2$，$q_1 := b/\beta$，$\alpha_1 := q_1^T A q_1$，
$r := Aq_1 - \alpha_1 q_1$，$\beta_1 := \|r\|_2$，$\xi_0 := 0$。

(3) 如果 $\beta_1 = 0$，则 $x := \frac{1}{\alpha_1} b$（此时，$\alpha_1$ 必然不为零），转步(10)；否则

$q_2:=r/\beta_1$，$\bar{\delta}_2:=\beta_1$，$\varepsilon_2:=0$，

$\gamma_1:=(\alpha_1^2+\beta_1^2)^{1/2}$，$c_1:=\alpha_1/\gamma_1$，$s_1:=\beta_1/\gamma_1$，

$\zeta_1:=\beta/\gamma_1$，$w_1:=c_1q_1+s_1q_2$，

$\tilde{w}_2:=s_1q_1-c_1q_2$，$\tilde{x}_1:=\zeta_1w_1$，$\Delta_0:=\beta$。

(4) 如果 $c_1\neq0$，则 $\Delta_1:=\dfrac{\beta\beta_1}{|\alpha_1|}$，$i:=2$；否则 $i:=2$。

(5) $\alpha_i:=q_i^{\mathrm{T}}Aq_i$，$r:=Aq_i-\alpha_iq_i-\beta_{i-1}q_{i-1}$，

$\beta_i:=\|r\|_2$，

$\delta_i:=\bar{\delta}_ic_{i-1}+\alpha_is_{i-1}$，

$\bar{\gamma}_i:=\bar{\delta}_is_{i-1}-\alpha_ic_{i-1}$，

$\varepsilon_{i+1}:=\beta_is_{i-1}$，

$\bar{\delta}_{i+1}:=-\beta_ic_{i-1}$，

$\gamma_i:=(\bar{\gamma}_i^2+\beta_i^2)^{1/2}$，

$c_i:=\bar{\gamma}_i/\gamma_i$，$s_i:=\beta_i/\gamma_i$。

(6) 如果 $c_i=0$，则转步(8)；否则

$$\Delta_i:=\begin{cases}\Delta_{i-1}|c_{i-1}s_i/c_i|, & c_{i-1}\neq0,\\ \Delta_{i-2}|c_{i-2}s_{i-1}s_i/c_i|, & c_{i-1}=0.\end{cases}$$

(7) 如果 $\Delta_i<\varepsilon$，则转步(9)；否则进行下一步。

(8) $q_{i+1}:=r/\beta_i$（此时 $\beta_i\neq0$，否则有 $c_i\neq0$而 $\Delta_i=0$），

$w_i:=c_i\tilde{w}_i+s_iq_{i+1}$，

$\tilde{w}_{i+1}:=s_i\tilde{w}_i-c_iq_{i+1}$，

$\zeta_i:=-(\varepsilon_i\zeta_{i-2}+\delta_i\zeta_{i-1})/\gamma_i$，

$\tilde{x}_i:=\tilde{x}_{i-1}+\zeta_iw_i$，

$i:=i+1$，转步(5)。

(9) $\bar{\zeta}_i:=-(\varepsilon_i\zeta_{i-2}+\delta_i\zeta_{i-1})/\bar{\gamma}_i$，

$x:=\tilde{x}_{i-1}+\bar{\zeta}_i\tilde{w}_i$。

(10) 输出 x，结束。

这一算法始终保持 A 不变，而且只涉及到 A 与向量的乘积，

因此可充分利用A的稀疏性，节约大量的内存，并有利于并行计算。此外，实际计算的经验还表明，常在迭代次数远小于A的阶数时，就可给出A的较好的近似解。因而，这一算法是求解大型稀疏对称线性方程组的一种有效的方法。

§5 求解非对称线性方程组的广义极小剩余法

这一节我们来介绍 Saad 和 Schultz(1986)给出的求解非对称线性方程组的广义极小剩余法(参见文献[60])。

设 $A\in\mathbb{R}^{n\times n}$ 非奇异，$b\in\mathbb{R}^n$ 非零。考虑线性方程组

$$Ax=b. \tag{5.1}$$

基于求解对称线性方程组的 Lanczos 方法的基本思想，我们也可按照如下的方式构造(5.1)的近似解 x_k：

(1) 任取一初始向量 x_0，并计算 $r_0=b-Ax_0$；

(2) 计算 Krylov 子空间$\mathscr{K}(A,r_0,k)$的一组标准正交基 $q_1,\cdots,q_k$；

(3) 构造

$$x_k=x_0+Q_ky_k,$$

其中 $Q_k=[q_1,\cdots,q_k]$，y_k 是方程组

$$Q_k^{\mathrm{T}}AQ_ky_k=Q_k^{\mathrm{T}}r_0$$

之解。

实现这一计算方案的关键在于如何实现Krylov子空间$\mathscr{K}(A,r_0,k)$之标准正交基的计算。类比于对称的情形，自然想到通过直接计算A的上 Hessenberg 分解来实现这一计算。

设A的上 Hessenberg 分解为

$$Q^{\mathrm{T}}AQ=H,$$

其中 $Q=[q_1,\cdots,q_n]$为正交矩阵，$H=[h_{ij}]$为上 Hessenberg阵。比较

$$AQ=QH$$

两边的第 j 列，可得

$$Aq_j = \sum_{i=1}^{j+1} h_{ij} q_i. \tag{5.2}$$

利用 Q 的正交性，可得

$$h_{ij} = q_i^{\mathrm{T}} A q_j, \quad i = 1, 2, \cdots, j, \tag{5.3}$$

$$h_{j+1,j} = \left\| Aq_j - \sum_{i=1}^{j} h_{ij} q_i \right\|_2, \tag{5.4}$$

$$h_{j+1,j} q_{j+1} = Aq_j - \sum_{i=1}^{j} h_{ij} q_i. \tag{5.5}$$

反过来，利用(5.3)—(5.5)，从任一给定的单位向量 q_1 出发，就可逐步产生 h_{ij} 和 q_j，这就是所谓的 **Arnoldi 方法**。类似于定理 1.1 的证明可证：这样产生的 $q_1, \cdots, q_j$ 正好是 Krylov 子空间 $\mathscr{K}(A, q_1, j)$ 的一组标准正交基，且

$$Q_j^{\mathrm{T}} A Q_j = H_j,$$

其中 $Q_j = [q_1, \cdots, q_j]$，H_j 是由迭代产生的 h_{ij} 所组成的 $j \times j$ 上 Hessenberg 阵；迭代终止（即出现 $h_{j+1,j} = 0$）的充分必要条件是 $j = \dim(\mathscr{K}(A, q_1, n))$。详见习题 6 。

利用 Arnoldi 方法的这些性质，并按照我们从求解对称线性方程组的 Lanczos方法所得到的基本思路，就可设计出求解(5.1)的如下算法：

算法5.1

(1) 输入 A 和初始向量 x_0。

(2) $r_0 := b - Ax_0$，$q_1 := r_0 / \|r_0\|_2$，$j := 1$。

(3) $h_{ij} := q_i^{\mathrm{T}} A q_j$，$i = 1, 2, \cdots, j$，

$$\hat{q}_{j+1} := Aq_j - \sum_{i=1}^{j} h_{ij} q_i,$$

$$h_{j+1,j} := \|\hat{q}_{j+1}\|_2.$$

(4) 如果 $h_{j+1,j} = 0$，则转步(5)；否则

$$q_{j+1} := \hat{q}_{j+1}/h_{j+1,j}, \qquad j := j+1,$$

转步(3)。

(5) 求解 $H_j y_j = \|r_0\| e_1$ 得 y_j,

$$x := x_0 + Q_j y_j,$$

输出 x_j，结束。

这一算法就是所谓求解线性方程组的 **Arnoldi 方法**。容易证明，由算法 5.1 得到的 x_j 满足

$$\|Ax_j - b\|_2 = h_{j+1,j} |e_j^T y_j| = 0,$$

即在没有误差出现的条件下，算法 5.1 在有限步将给出(5.1)的精确解。在误差出现时，当然亦可用 $|h_{j+1,j} e_j^T y_j|$ 是否很小来判断迭代是否终止。

但是，这里需要指出的是，当 j 很大时，由于整个计算过程需要保存所有的向量 $q_1, \cdots, q_j$，因而需占用大量的内存，以致于使得所处理的问题的规模受到很大的限制。

为了克服 Arnoldi 方法的这一缺点，通常使用时，先选择一个适当的正整数 m(一般不宜太大)，计算到 $j=m$ 时，就去求 x_m；然后再以 x_m 作为初始向量重新开始；这样周而复始直到求出所需精度的近似解为止。这就是所谓的 **循环 Arnoldi 方法**。这样做，虽然解决了存储问题，但又会遇到方程组 $H_m y_m = \|r_0\|_2 e_1$ 不相容的情形，致使迭代半途而废。

针对循环 Arnoldi 方法的这一缺点，Saad 和 Schultz(1986) 提出了广义极小剩余法（参见文献[60]）。他们的基本思想是：对于选定的初值 x_0，第 k 步是求剩余泛函

$$\psi(x) = \|b - Ax\|_2$$

在 k 维超平面

$$x_0 + \mathscr{K}(A, r_0, k)$$

上的极小点 x_k，其中 $r_0 = b - Ax_0$，即

$$x_k = x_0 + z_k,$$

z_k 是最小二乘问题

$$\min_{z\in\mathscr{K}(A,r_0,k)}\|b-A(x_0+z)\|_2=\min_{z\in\mathscr{K}(A,r_0,k)}\|r_0-Az\|_2 \quad (5.6)$$

的解。

现在假定我们从 $q_1=r_0/\|r_0\|_2$ 出发，利用 Arnoldi 方法已求出正交向量组 $q_1,\cdots,q_{k+1}$ 和 h_{ij}。则有

$$AQ_k=Q_{k+1}\bar{H}_k, \quad (5.7)$$

其中 $Q_i=[q_1,\cdots,q_i]$，$i=k,k+1$；$\bar{H}_k\in\mathbf{R}^{(k+1)\times k}$ 为

$$\bar{H}_k=\begin{bmatrix} h_{11} & \cdots & h_{1,k-1} & h_{1k} \\ h_{21} & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ 0 & & h_{k,k-1} & h_{kk} \\ & & & h_{k+1,k} \end{bmatrix}. \quad (5.8)$$

利用(5.7)可得

$$\begin{aligned}\min_{z\in\mathscr{K}(A,r_0,k)}\|r_0-Az\|_2&=\min\{\|r_0-AQ_ky\|_2:\ y\in\mathbf{R}^k\}\\ &=\min\{\|r_0-Q_{k+1}\bar{H}_ky\|_2:\ y\in\mathbf{R}^k\}\\ &=\min\{\|\|r_0\|_2e_1-\bar{H}_ky\|_2:\ y\in\mathbf{R}^k\},\end{aligned} \quad (5.9)$$

其中最后一个等式用到了 $r_0=\|r_0\|q_1$ 和 $Q_{k+1}^{\mathrm{T}}Q_{k+1}=I_{k+1}$。由于 $\bar{H}_k$ 的特殊形式，(5.9)是容易求解的。这样，只需在算法5.1中将求解方程组 $H_ky_k=\|r_0\|e_1$ 换成求最小二乘问题(5.9)就得到了广义极小剩余法，简称 GMRES 算法。而实际上常用的是循环 GMRES算法，即迭代 m 步后再重新开始，习惯上称作GMRES(m)算法，可简述如下：

算法5.2(GMRES(m)算法)

(1) 输入 A,b，初始向量 x_0，最大迭代次数 m 及精度要求 ε。

(2) $r_0:=b-Ax_0$，$q_1:=r_0/\|r_0\|_2$，$j:=1$。

(3) $h_{ij}:=q_i^{\mathrm{T}}Aq_j$，$i=1,2,\cdots,j$，

$$\hat{q}_{j+1}:=Aq_j-\sum_{i=1}^{j}h_{ij}q_i,$$

$$h_{j+1,j}:=\|\hat{q}_{j+1}\|_2.$$

（4）如果 $h_{j+1,j}<\varepsilon$ 或 $j=m$，则转步(5)；否则

$$q_{j+1}:=\hat{q}_{j+1}/h_{j+1,j},\qquad j:=j+1,$$

转步(3)。

（5）求解最小二乘问题

$$\min\{\|\tilde{H}_j y-\|r_0\|e_1\|_2 : y\in\mathbb{R}^j\}$$

得 y_j，

$$x_j:=x_0+Q_j y_j.$$

（6）如果 $\|Ax_j-b\|_2<\varepsilon$，则输出 x_j，结束；否则 $x_0:=x_j$，转步(2)。

显然，算法5.2不会发生中断，弥补了循环Arnoldi方法的不足；而且与第五章所介绍的广义共轭剩余法（即GCR）相比，有节约存储量和运算量少等优点。因此，目前人们认为这一算法是求解大型稀疏非对称线性方程组的最有效方法之一。

至于算法5.2中的 m 取多大为最好，现在还没有理论上的结果。在理论上仅可以保证，在系数矩阵 A 有正定的对称部分 $\left(\text{即}\frac{1}{2}(A^T+A)\text{正定}\right)$ 时，对任意的 m，算法5.2总是收敛的，即随着循环次数的增加剩余向量趋向于零（详见文献[60]）。但对一般情形，并非对任意的 m 总是收敛的。例如，对线性方程组

$$\begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\begin{bmatrix}1\\ 1\end{bmatrix}.$$

取 $x_0=0$，利用GMRES(1)求解，不论循环多少次，总有 $x_j=0$，而 $\|Ax_j-b\|_2=\sqrt{2}$，永远不会得到原方程组的解。但如果用GMRES求解，两步就可得到方程组的精确解。

另外需指出的是，利用Arnoldi方法求Krylov子空间的标准正交基的过程，实质上就是Gram-Schmidt正交化过程。因此，在误差出现的情况下，算法的数值稳定性较差。基于这种考虑，

Wolker(1988)提出了用 Householder 变换求 Krylov 子空间的标准正交基的 GMRES 算法。

假设 $v_1 \in \mathbf{R}^n$，$\|v_1\|_2 = 1$。先取 Householder 变换 P_1，使 $P_1 v_1 = e_1$。然后对 $m = 1, 2, \cdots$ 迭代执行：

(1) $v_m := P_1 \cdots P_m e_m$。

(2) 如果 $\operatorname{rank}[v_1, Av_1, \cdots, Av_m] = m$，则结束；否则，选取 Householder 矩阵 P_{m+1} 使

$$P_{m+1} P_m \cdots P_1 [v_1, Av_1, \cdots, Av_m]$$

为上三角矩阵。

容易验证，按照上述迭代产生的向量 $v_1, \cdots, v_m$ 是 Krylov 子空间 $\mathscr{K}(A, v_0, m)$ 的一组标准正交基。

现在假定对给定的初始向量 x_0，从 $v_1 = r_0 / \|r_0\|_2 (r_0 = b - Ax_0)$ 出发，利用上述方法产生了向量 $v_1, \cdots, v_m$ 和 Householder 矩阵 $P_1, \cdots, P_{m+1}$。令

$$P_{m+1} \cdots P_1 [Av_1, \cdots, Av_m] = \underset{m}{\begin{bmatrix} H_m \\ 0 \end{bmatrix}} \begin{matrix} m+1 \\ n-m+1 \end{matrix},$$

则 H_m 是 $(m+1) \times m$ 的上 Hessenberg 阵，且有

$$\min_{z \in \mathscr{K}(A, r_0, m)} \|b - A(x_0 + z)\|_2 = \min\{\| \|r_0\|_2 e_1 - H_m y \|_2 : y \in \mathbf{R}^m\}.$$

由此不难给出利用 Householder 变换执行的 GMRES 算法，建议读者作为练习给出其详情细节。

习　题

1. 设 T 是 n 阶不可约实对称三对角矩阵。试证：T 的任何特征向量的第一个分量不为零。

2. 证明：对任意的自然数 $k \leqslant j \leqslant n$，和非零向量 $v \in \mathscr{K}(A, q_1, j) \cap \operatorname{span}\{x_k, \cdots, x_n\}$，有

$$\mu_k \leqslant \rho(v; A) + \sum_{i=1}^{k-1} (\lambda_n - \mu_i) \sin^2 \theta_i;$$

由此并利用第二节的有关引理及技巧证明 Kaniel 关于 $\mu_k-\lambda_k$ 的收敛性估计：

$$0\leqslant\mu_k-\lambda_k\leqslant\frac{(\lambda_n-\lambda_k)}{|C_{j-k}(1+2r)|^2}\sum_{i=k+1}^{n}\left(\frac{\xi_i}{\xi_k}\right)^2\prod_{i=1}^{k-1}\left(\frac{\lambda_i-\lambda_n}{\lambda_i-\lambda_k}\right)^2$$

$$+\sum_{i=1}^{k-1}(\lambda_n-\lambda_i)\sin^2\theta_i.$$

其中 $q_1,\xi_i,x_i,\mu_i,\lambda_i,\theta_i$ 和 $C_{j-k}(t)$ 如第二节所定义。

3．试给出一种不用正交变换二对角化一个给定矩阵 $A\in\mathbb{R}^{m\times n}$ 的方法。并利用你的方法设计出适用于大型稀疏矩阵的奇异值分解和最小二乘问题的数值方法。

4．证明：如果 $A\in S\mathbb{R}^{n\times n}$ 有重特征值，则 Lanczos迭代必然在 $k<n$ 时终止（即出现 $\beta_k=0$)。

5．设 $A\in S\mathbb{R}^{n\times n}$，$q_1\in\mathbb{R}^n$，$\|q_1\|_2=1$，并假定 m 是以初始向量 q_1 进行 Lanczos 迭代的终止指标（即 $\beta_m=0$，$\beta_i\neq0,1\leqslant i\leqslant m-1$)。证明：$m$ 是包含 q_1 的 A 的不变子空间的最小维数。

6．证明 Arnoldi 方法产生的 q_i 和 h_{ij} 满足：

(1) 正数 m 满足 $h_{j+1,j}\neq0$，$1\leqslant j<m$，而 $h_{m+1,m}=0$ 的充分必要条件是 $m=\dim(\mathscr{K}(A,q_1,n))$；

(2) 对任意的 $1\leqslant j\leqslant m$ 有：

(a) $Q_j^{\mathrm T}Q_j=I_j$，　　(b) $Q_j^{\mathrm T}AQ_j=H_j$，

(c) $\mathscr{K}(A,q_1,j)=\operatorname{span}\{q_1,\cdots,q_j\}$.

7．设线性方程组的系数矩阵为 $A=[e_2,\cdots,e_n,e_1]\in\mathbb{R}^{n\times n}$，$b=e_1$。对任意的 $1\leqslant m\leqslant n$，利用 Arnoldi 方法求列正交矩阵 $Q_m=[q_1,\cdots,q_m]\in\mathbb{R}^{n\times m}$ 和上 Hessenberg 矩阵 H_m，使得

$$Q_m^{\mathrm T}AQ_m=H_m\ 且\ q_1=b=e_1;$$

并证明：对任意的 $m<n$，方程组 $H_my=e_1$ 总是不相容的。

8．给出利用 Householder 方法求 Krylov子空间的标准正交

基的广义极小剩余法的详细算法。

9．证明恒等式(5.7)。并由此证明算法 5.1 产生的x_j满足

$$\|Ax_j-b\|_2=h_{j+1,j}|e_j^{\mathrm{T}}y_j|.$$

10．设计一个数值稳定的求解最小二乘问题（5.9）的算法。

第十章　求解Jacobi 矩阵特征值反问题的数值方法

§1　基本问题和定性理论

所谓Jacobi 矩阵是指形如

$$T=\begin{bmatrix}\alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \ddots & & 0 \\ & \ddots & \ddots & \ddots & \\ & 0 & \ddots & \alpha_{n-1} & \beta_n \\ & & & \beta_n & \alpha_n\end{bmatrix}$$

的实对称三对角矩阵，其中 $\beta_i>0$，$i=2,\cdots,n$。对于给定的 n 阶 Jacobi 矩阵 T，我们用 $\hat{T}$ 来表示 T 划去第一列和第一行之后所得到的 $n-1$ 阶主子阵，通常称作 T 的右下角的 $n-1$ 阶主子阵。

Jacobi 矩阵特征值反问题就是根据已知的特征值和特征向量的某些信息求 Jacobi 矩阵的元素。这类问题产生于地球物理、振动力学等应用学科，由于实际问题的差异而提出的问题也不尽相同，但其中最基本的问题是：

问题 1　给定满足如下条件

$$\lambda_1<\mu_1<\lambda_2<\mu_2<\cdots<\mu_{n-1}<\lambda_n \tag{1,1}$$

的 $2n-1$ 个实数，求一个 n 阶 Jacobi 矩阵 T，使得 T 和它的右下角的 $n-1$ 阶主子阵 $\hat{T}$ 的特征值分别为 $\lambda_1,\cdots,\lambda_n$ 和 $\mu_1,\cdots,\mu_{n-1}$。

这类问题是由 Hochstadt 于1967年首先提出的，经过近二十多年来的深入研究，现在已达到实用阶段。理论上，已证问题 1 的解存在且唯一，并连续地依赖于给定的数据；数值方法上，已得到几种快速稳定的方法。这一节，先来证明问题 1 之解的存在

唯一性，下一节介绍有关的数值方法。

定理1.1 设T是n阶 Jacobi 矩阵。则T是问题 1 之解的充分必要条件是T有分解

$$T = Q\Lambda Q^{\mathrm{T}}, \tag{1.2}$$

其中$\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$，$Q = [q_{ij}]$正交，而且$Q$的第一行元素满足

$$q_{1j}^2 = \prod_{i=1}^{n-1}(\lambda_j - \mu_i) \Big/ \prod_{\substack{i=1 \\ i \neq j}}^{n}(\lambda_j - \lambda_i), \quad j = 1, 2, \cdots, n. \tag{1.3}$$

证明 必要性 若T是问题 1 之解，则由$\lambda_1, \cdots, \lambda_n$是$T$的特征值，故存在正交矩阵$Q = [q_{ij}]$使得(1.2)成立。又$T$的右下角的$n-1$阶主子阵$\hat{T}$的特征值是$\mu_1, \cdots, \mu_{n-1}$，则由第九章引理3.1知(1.3)成立。

充分性 设T有分解(1.2)并满足(1.3)。记

$$p_n(\lambda) = \det(\lambda I - T), \quad p_{n-1}(\lambda) = \det(\lambda I - \hat{T}).$$

由第九章引理3.1知

$$q_{1j}^2 = p_{n-1}(\lambda_j)/p_n'(\lambda_j), \quad j = 1, 2, \cdots, n. \tag{1.4}$$

由T满足(1.2)知$p_n(\lambda) = \prod_{j=1}^{n}(\lambda - \lambda_j)$，从而由(1.3)和(1.4)可得

$$p_{n-1}(\lambda_j) = \prod_{i=1}^{n-1}(\lambda_j - \mu_i), \quad j = 1, 2, \cdots, n. \tag{1.5}$$

注意到$p_{n-1}(\lambda)$和$\prod_{i=1}^{n-1}(\lambda - \mu_i)$都是$\lambda$的$n-1$次首一多项式，$\lambda_i$互不相同，即知(1.5)蕴含着

$$p_{n-1}(\lambda) \equiv \prod_{i=1}^{n-1}(\lambda - \mu_i).$$

从而$\hat{T}$的特征值是$\mu_1, \cdots, \mu_{n-1}$，即T是问题 1 之解。

定理1.2 问题 1 有且仅有一个解。

证明 *存在性* 由给定的 $2n-1$ 个数定义两个多项式：

$$p_n(\lambda)=\prod_{i=1}^{n}(\lambda-\lambda_i) \text{ 和 } p_{n-1}(\lambda)=\prod_{i=1}^{n-1}(\lambda-\mu_i).$$

下证存在实数 α_1，正数 γ_2 和一个 $n-2$ 次首一多项式 $p_{n-2}(\lambda)$ 满足

$$p_n(\lambda)=(\lambda-\alpha_1)p_{n-1}(\lambda)-\gamma_2 p_{n-2}(\lambda), \tag{1.6}$$

并且 $p_{n-2}(\lambda)$ 的零点严格分隔 $p_{n-1}(\lambda)$ 的零点。

设

$$p_n(\lambda)=\lambda^n+a_{n-1}\lambda^{n-1}+\cdots+a_1\lambda+a_0,$$

$$p_{n-1}(\lambda)=\lambda^{n-1}+b_{n-2}\lambda^{n-2}+\cdots+b_1\lambda+b_0,$$

$$p_{n-2}(\lambda)=\lambda^{n-2}+c_{n-3}\lambda^{n-3}+\cdots+c_1\lambda+c_0,$$

并代入(1.6)，比较两边同次幂的系数，可得

$$\alpha_1=b_{n-2}-a_{n-1},$$

$$\gamma_2=b_{n-3}-\alpha_1 b_{n-2}-a_{n-2},$$

$$c_j=(b_{j-1}-\alpha_1 b_j-a_j)/\gamma_2,\quad j=n-3,\cdots,1,$$

$$c_0=-(a_0+\alpha_1 b_0)/\gamma_2.$$

只要证明了 $\gamma_2>0$，则上述公式就有意义，从而也就找到了所需的数 α_1,γ_2 和多项式 p_{n-2}。

由 p_n 和 p_{n-1} 的定义，可得

$$a_{n-1}=-\sum_{i=1}^{n}\lambda_i,\qquad a_{n-2}=\sum_{j=1}^{n-1}\lambda_{j+1}\sum_{k=1}^{j}\lambda_k,$$

$$b_{n-2}=-\sum_{i=1}^{n-1}\mu_i,\qquad b_{n-3}=\sum_{j=1}^{n-2}\mu_j\sum_{k=j+1}^{n-1}\mu_k.$$

于是，有

$$\alpha_1=\sum_{j=1}^{n}\lambda_j-\sum_{j=1}^{n-1}\mu_j,$$

$$\gamma_2=\sum_{j=1}^{n-1}(\lambda_{j+1}-\mu_j)\sum_{k=1}^{j}(\mu_k-\lambda_k).$$

从已知的 $2n-1$ 个数满足条件(1.1)立即知道 $\gamma_2>0$。

在(1.6)中令 $\lambda=\mu_k$ 即得

$$-\gamma_2 p_{n-2}(\mu_k)=\prod_{j=1}^{n}(\mu_k-\lambda_j),\quad k=1,2,\cdots,n-1.$$

由此即知 $p_{n-2}(\mu_k)$的符号为$(-1)^{n-k+1}$。所以在每个区间(μ_k,μ_{k+1})内必有 $p_{n-2}(\lambda)$的一个零点 ν_k。又 $p_{n-2}(\lambda)$至多有$n-2$个互不相同的零点，因此，$p_{n-2}(\lambda)$的零点正好严格分隔 $p_{n-1}(\lambda)$的零点，即

$$\mu_1<\nu_1<\mu_2<\cdots<\nu_k<\mu_k<\cdots<\nu_{n-2}<\mu_{n-1}.$$

重复前面的推理，由 p_{n-1} 和 p_{n-2} 又可找到实数 α_2，正数 γ_3 和 $n-3$ 次首一多项式 $p_{n-3}(\lambda)$满足

$$p_{n-1}(\lambda)=(\lambda-\alpha_2)p_{n-2}(\lambda)-\gamma_3 p_{n-3}(\lambda),$$

而且 $p_{n-3}(\lambda)$的零点严格分隔 $p_{n-2}(\lambda)$的零点。

如此进行 $n-1$ 步，就可找到$n-1$个实数 $\alpha_1,\alpha_2,\cdots,\alpha_{n-1}$，$n-1$个正数 $\gamma_2,\gamma_3,\cdots,\gamma_n$，以及$n-1$个首一多项式 $p_{n-2}(\lambda)$，$p_{n-3}(\lambda),\cdots,p_1(\lambda),p_0(\lambda)$。满足：

(i) $p_{n-k+1}(\lambda)=(\lambda-\alpha_k)p_{n-k}(\lambda)-\gamma_{k+1}p_{n-k-1}(\lambda),k=1,\cdots,n-1$；

(ii) $p_k(\lambda)$是k次多项式($p_0(\lambda)\equiv 1$)；

(iii) $p_{k-1}(\lambda)$的零点严格分隔 $p_k(\lambda)$的零点。

令

$$\alpha_n=p_1(0),$$

$$\beta_k=\sqrt{\gamma_k},\quad k=2,\cdots,n,$$

$$T=\begin{bmatrix}\alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \ddots & & 0\\ & \ddots & \ddots & \ddots & \\ 0 & & \ddots & \alpha_{n-1} & \beta_n\\ & & & \beta_n & \alpha_n\end{bmatrix}.$$

则易证这样构造出的T之特征多项式必为p_n，而其右下角的$n-1$

阶主子阵$\tilde{T}$的特征多项式正好是p_{n-1}。从而T是问题1的一个解。

唯一性 现假定T和$\tilde{T}$是问题1的两个解。由定理1.1知，T和$\tilde{T}$有分解

$$T=Q\Lambda Q^{\mathrm{T}} \text{ 和 } \tilde{T}=\tilde{Q}\Lambda\tilde{Q}^{\mathrm{T}}, \tag{1.7}$$

其中$\Lambda=\operatorname{diag}(\lambda_1,\cdots,\lambda_n)$，$Q=[q_{ij}]$和$\tilde{Q}=[\tilde{q}_{ij}]$正交且它们的第一行都满足(1.3)。显然，我们可以通过调整Q和$\tilde{Q}$每列元素的符号，使得它们的第一行元素均为正数，这样便有

$$q_{1j}=\tilde{q}_{1j}, \quad j=1,2,\cdots,n. \tag{1.8}$$

从(1.8)和(1.7)利用归纳法容易推出$Q=\tilde{Q}$，$T=\tilde{T}$（参照第七章定理2.2的证明）。

定理1.3 设T和$\tilde{T}$分别是问题1关于给定数据

$$\lambda_1<\mu_1<\lambda_2<\cdots<\mu_{n-1}<\lambda_n$$

和

$$\tilde{\lambda}_1<\tilde{\mu}_1<\tilde{\lambda}_2<\cdots<\tilde{\mu}_{n-1}<\tilde{\lambda}$$

所对应的解。则存在正数K和δ使得，只要

$$\sum_{j=1}^{n}(\lambda_j-\tilde{\lambda}_j)^2+\sum_{j=1}^{n-1}(\mu_i-\tilde{\mu}_i)^2<\delta,$$

就有

$$\|T-\tilde{T}\|_F\leqslant K\left(\sum_{j=1}^{n}(\lambda_j-\tilde{\lambda}_j)^2+\sum_{j=1}^{n-1}(\mu_j-\tilde{\mu}_j)^2\right)^{1/2}. \tag{1.9}$$

这一定理的证明较繁，这里不再给出，有兴趣的读者可以参看文献[39]或[72]。

(1.9)表明，将问题1之解看作给定数据的函数的话，它是局部Lipschitz连续的。

§2 数值方法

2.1 Lanczos方法

根据定理1.1和1.2，并利用第九章所介绍的Lanczos迭代，我们可以按如下步骤来计算问题1的解：

(1) 用(1.3)求出一个单位向量 q；

(2) 对 $\Lambda=\mathrm{diag}(\lambda_1,\cdots,\lambda_n)$和 q 应用Lanczos 迭代求出Jacobi矩阵T。

注意到 Λ 的简单形式，可设计算法如下：

算法2.1

(1) 输入 $\lambda_1,\cdots,\lambda_n$ 和 $\mu_1,\cdots,\mu_{n-1}$。

(2) $\omega_i:=\left[\prod_{j=1}^{n-1}(\mu_j-\lambda_i)\Big/\prod_{\substack{j=1\\ j\neq i}}^{n}(\lambda_j-\lambda_i)\right]^{1/2}\quad(i=1,\cdots,n)$,

$$a_1:=\sum_{i=1}^{n}\lambda_i\omega_i^2,\quad j:=1.$$

(3) 如果 $j=1$，则

$$\nu_i:=\omega_i(\lambda_i-a_1)\ (i=1,2,\cdots,n);$$

否则，

$$\nu_i:=(\lambda_i-a_j)\omega_i-\beta_j\gamma_i\ (i=1,2,\cdots,n).$$

(4) $\beta_{j+1}:=\left(\sum_{i=1}^{n}\nu_i^2\right)^{1/2}$,

$\gamma_i:=\omega_i\ (i=1,2,\cdots,n)$,

$\omega_i:=\nu_i/\beta_{j+1}\ (i=1,2,\cdots,n)$.

$a_{j+1}:=\sum_{i=1}^{n}\lambda_i\omega_i^2$.

(5) 如果 $j<n-1$，则 $j:=j+1$转(3)；否则输出 $a_1,\cdots,a_n$和$\beta_2,\cdots,\beta_n$，结束。

这一算法所需运算量是 $O(n^2)$。但由于 Lanczos 方法产生的 Lanczos 向量很快失去正交性，因而这一算法的数值稳定性较差。为了保证数值稳定性，必须使用再正交化技巧；然而，这样做导致所需运算量增到 $O(n^3)$。

2.2 正交约化法

前面介绍的 Lanczos 方法，虽然简单易行，但为了保证数值

稳定性，必须进行再正交化，这样运算量又太大。针对这一问题，经过长期研究，现在终于设计出另一种十分漂亮的算法——正交约化法。这种方法既数值稳定性好，又仅需 $O(n^2)$ 的运算量即可完成。

对给定的数据，用(1.3)计算出单位向量 q 之后，构造如下的对角加边矩阵

$$A=\begin{bmatrix} \alpha_0 & q^T \\ q & \Lambda \end{bmatrix} \in S\mathbb{R}^{(n+1)\times(n+1)}, \tag{2.1}$$

其中 $\Lambda=\text{diag}(\lambda_1,\cdots,\lambda_n)$，$\alpha_0$ 是哑元，可随意指定。如果能够找到一个 n 阶正交矩阵 Q，使得

$$\begin{bmatrix} 1 & 0 \\ 0 & Q \end{bmatrix}\begin{bmatrix} \alpha_0 & q^T \\ q & \Lambda \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & Q^T \end{bmatrix}=\begin{bmatrix} \alpha_0 & \beta_1 e_1^T \\ \beta_1 e_1 & T \end{bmatrix}, \tag{2.2}$$

其中 T 是实对称三对角矩阵，则有

$$Q\Lambda Q^T=T, \quad \beta_1 Q^T e_1=q, \quad \beta_1=\pm\|q\|_2.$$

再注意到，对任意的不可约实对称三对角矩阵 T，都可构造一个对角元素为 1 或 -1 的对角阵 D 使得 $D^{-1}TD$ 的次对角元素均为正数，我们再将(2.2)中所得到的 T 的所有次对角元素均取绝对值就得到了问题 1 的解(定理1.1)。

这样一来，求解问题 1 就转化为如何求形如(2.2)的分解，即如何利用第一列均为单位向量 $e_1^{(n+1)}$ 的正交变换将 A 正交约化为对称三对角矩阵。从第七章约化一般实对称矩阵为对称三对角矩阵的标准方法可知，可用 Householder 方法来完成这一任务，但这样做所需的运算量为 $O(n^3)$。注意到(2.1)所定义的 A 的特殊性，利用 Givens 变换来约化，则可大大减少约化所需的运算量。下面介绍这方面十分漂亮的两种方法。

1. 驱逐出境法

这一方法是受对称 QR 方法的启发而得到的，其运算量与下面将要介绍的 Rutishauser 方法完全相同，但这一方法更自然一

些。具体执行过程不难从下面的矩阵图示中明白。

$$A=\begin{bmatrix}\times&\times&\times&\times&\times\\\times&\times&0&0&0\\\times&0&\times&0&0\\\times&0&0&\times&0\\\times&0&0&0&\times\end{bmatrix}\xrightarrow{(4,5)}\begin{bmatrix}\times&\times&\times&\times&0\\\times&\times&0&0&0\\\times&0&\times&0&0\\\times&0&0&\times&+\\0&0&0&+&\times\end{bmatrix}$$

$$\xrightarrow{(3,4)}\begin{bmatrix}\times&\times&\times&0&0\\\times&\times&0&0&0\\\times&0&\times&+&+\\0&0&+&\times&\times\\0&0&+&\times&\times\end{bmatrix}\xrightarrow{(4,5)}\begin{bmatrix}\times&\times&\times&0&0\\\times&\times&0&0&0\\\times&0&\times&\times&0\\0&0&\times&\times&\times\\0&0&0&\times&\times\end{bmatrix}$$

$$\xrightarrow{(2,3)}\begin{bmatrix}\times&\times&0&0&0\\\times&\times&+&+&0\\0&+&\times&\times&0\\0&+&\times&\times&\times\\0&0&0&\times&\times\end{bmatrix}\xrightarrow{(3,4)}\begin{bmatrix}\times&\times&0&0&0\\\times&\times&\times&0&0\\0&\times&\times&\times&+\\0&0&\times&\times&\times\\0&0&+&\times&\times\end{bmatrix}$$

$$\xrightarrow{(4,5)}\begin{bmatrix}\times&\times&0&0&0\\\times&\times&\times&0&0\\0&\times&\times&\times&0\\0&0&\times&\times&\times\\0&0&0&\times&\times\end{bmatrix},$$

其中“$\times$”表示可能有的非零元素，“$+$”表示变换后可能新增加的非零元素，“$\xrightarrow{(i,j)}$”表示对箭尾所指的矩阵进行(i,j)坐标平面内的正交相似变换后变成箭头所指的矩阵。

上述过程，每消去一个边上的非零元素，就会出现一个不希望有的新的非零元，然后再将这一非零元素逐步消除。例如，当我们消去(1,3)和(3,1)位置上的非零元素之后，就会在(2,4)和

(4,2)位置上出现一个不希望有的非零元素；然后，我们就接着消去这一非零元素，于是又在(3,5)和(5,3)位置上又出现了一个新的不希望有的非零元素；紧接着再消去这一非零元素，就得到了我们所需的形式。因而，我们形象地称这一约化法为驱逐出境法。

一般地，如果从消去$(1,n+1)$和$(n+1,1)$位置的非零元素出发，已进行了k步，第$k+1$步是先消去$(n-k,1)$和$(1,n-k)$位置上的非零元素，此时就会在$(n-k+1,n-k-1)$和$(n-k-1,n-k+1)$位置上出现一个新的不希望有的非零元素；接着连续进行$(n-k,n-k+1)$，$(n-k+1,n-k+2)$，$\cdots$，$(n,n+1)$平面内的旋转变换，就可将这一不希望有的非零元素逐步从矩阵的内部驱赶到矩阵之外。

综上所述可得如下算法。

算法2.2

(1) 输入$\lambda_1,\cdots,\lambda_n$和$\mu_1,\cdots,\mu_{n-1}$。

(2) $\omega_i:=\left[\prod_{j=1}^{n-1}(\mu_j-\lambda_i)\Big/\prod_{\substack{j=1\\j\neq i}}^{n}(\lambda_j-\lambda_i)\right]^{\frac{1}{2}}\quad(i=1,\cdots,n)$,

$\alpha_i:=\lambda_i\ (i=1,2,\cdots,n)$,

$\beta_i:=0\ (i=2,\cdots,n)$,

$i:=n$.

(3) 确定$s=\sin\theta$，$c=\cos\theta$，使

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix}\begin{bmatrix} \omega_{i-1} \\ \omega_i \end{bmatrix}=\begin{bmatrix} * \\ 0 \end{bmatrix},$$

$$\omega_{i-1}:=c\omega_{i-1}+s\omega_i,$$

$$\begin{bmatrix} \alpha_{i-1} & \beta_i \\ \beta_i & \alpha_i \end{bmatrix}:=\begin{bmatrix} c & s \\ -s & c \end{bmatrix}\begin{bmatrix} \alpha_{i-1} & 0 \\ 0 & \alpha_i \end{bmatrix}\begin{bmatrix} c & -s \\ s & c \end{bmatrix}.$$

(4) 如果$i=n$，则$i:=i-1$，转步(3)；否则$k:=i$，进行下一步。

(5) $$\begin{bmatrix} \alpha \\ \beta_{k+1} \end{bmatrix} := \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} 0 \\ \beta_{k+1} \end{bmatrix},$$

确定 $s=\sin\theta$，$c=\cos\theta$，使

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} \beta_k \\ \alpha \end{bmatrix} = \begin{bmatrix} * \\ 0 \end{bmatrix},$$

$$\beta_k := c\beta_k + s\alpha,$$

$$\begin{bmatrix} \alpha_k & \beta_{k+1} \\ \beta_{k+1} & \alpha_{k+1} \end{bmatrix} := \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} \alpha_k & \beta_{k+1} \\ \beta_{k+1} & \alpha_{k+1} \end{bmatrix} \begin{bmatrix} c & -s \\ s & c \end{bmatrix}.$$

(6) 如果 $k<n-1$，则 $k:=k+1$，转步(5)；否则进行下一步。

(7) 如果 $i>2$，则 $i:=i-1$，转步(3)；否则

$$\beta_i := |\beta_i|, \quad i=2,\cdots,n,$$

输出 $\alpha_1,\cdots,\alpha_n$ 和 $\beta_2,\cdots,\beta_n$，结束。

不难算出这一算法的运算量是 $O(n^2)$。而且由于整个约化过程是用数值稳定的 Givens 变换进行的，所以这一算法是数值稳定的。数值试验的结果亦表明这一算法是快速有效的。

2. Rutishauser 方法

这一方法基本上类似于前面介绍的驱逐出境法，只是约化的次序不同，其约化过程不难从下面的图示明白。

$$A = \begin{bmatrix} \times & \times & \times & \times & \times \\ \times & \times & 0 & 0 & 0 \\ \times & 0 & \times & 0 & 0 \\ \times & 0 & 0 & \times & 0 \\ \times & 0 & 0 & 0 & \times \end{bmatrix} \xrightarrow{(2,3)} \begin{bmatrix} \times & \times & 0 & \times & \times \\ \times & \times & + & 0 & 0 \\ 0 & + & \times & 0 & 0 \\ \times & 0 & 0 & \times & 0 \\ \times & 0 & 0 & 0 & \times \end{bmatrix}$$

$$\xrightarrow{(2,4))} \begin{bmatrix} \times & \times & 0 & 0 & \times \\ \times & \times & \times & + & 0 \\ 0 & \times & \times & + & 0 \\ 0 & + & + & \times & 0 \\ \times & 0 & 0 & 0 & \times \end{bmatrix} \xrightarrow{(3,4)} \begin{bmatrix} \times & \times & 0 & 0 & \times \\ \times & \times & \times & 0 & 0 \\ 0 & \times & \times & \times & 0 \\ 0 & 0 & \times & \times & 0 \\ \times & 0 & 0 & 0 & \times \end{bmatrix}$$

$$\xrightarrow{(2,5)}\begin{bmatrix}\times&\times&0&0&0\\\times&\times&\times&0&+\\0&\times&\times&\times&+\\0&0&\times&\times&0\\0&+&+&0&\times\end{bmatrix}\xrightarrow{(3,5)}\begin{bmatrix}\times&\times&0&0&0\\\times&\times&\times&0&0\\0&\times&\times&\times&\times\\0&0&\times&\times&+\\0&0&\times&+&\times\end{bmatrix}$$

$$\xrightarrow{(4,5)}\begin{bmatrix}\times&\times&0&0&0\\\times&\times&\times&0&0\\0&\times&\times&\times&0\\0&0&\times&\times&\times\\0&0&0&\times&\times\end{bmatrix}.$$

一般地，如果我们已进行了若干步之后，已将 A 正交相似变换为如下形状的矩阵

$$\tilde{A}=\begin{bmatrix}J & B\\ B^{\mathrm{T}} & D\end{bmatrix}\begin{matrix}i\\ n-i+1\end{matrix},\qquad (2.3)$$

（列块宽度分别为 i，$n-i+1$）

其中 J 是实对称三对角矩阵，$D=\operatorname{diag}(\lambda_i,\cdots,\lambda_n)$，

$$B=\begin{bmatrix}\omega_i & \omega_{i+1} & \cdots & \omega_n\\ 0 & 0 & \cdots & 0\\ \cdots & \cdots & \cdots & \cdots\\ 0 & 0 & \cdots & 0\end{bmatrix},\quad \omega_i=e_i^{\mathrm{T}}q.$$

我们下一步是先进行 $(2,i+1)$ 平面内的正交相似变换将 $\tilde{A}$ 之 $(i+1,1)$ 和 $(1,i+1)$ 位置的非零元素 ω_i 化为零，这样就在 $(i+1,2)$，$(i+1,3)$，$(2,i+1)$ 和 $(3,i+1)$ 出现非零元素；然后连续进行 $(3,i+1)$，$(4,i+1)$，$\cdots$，$(i,i+1)$ 平面内的正交相似变换，就又将其变换为形如(2,3)的矩阵，不过此时 J 增加了一阶，而 B 和 D 都减少了一阶。

这一约化过程的详细算法建议读者作为练习将其总结出来。

§3 相关问题

这一节我们来介绍几类可用上节介绍的方法求解的问题。

3.1 秩1修改问题

问题2 给定 $2n$ 个实数

$$\lambda_1<\mu_1<\lambda_2<\mu_2<\cdots<\lambda_n<\mu_n,$$

求一个 n 阶 Jacobi 矩阵 T，使得 T 的特征值是 $\lambda_1,\cdots,\lambda_n$，而将 T 之(1,1)位置上的元素作适当修改之后得到的 Jacobi 阵 $\tilde{T}$ 就有特征值 $\mu_1,\mu_2,\cdots,\mu_n$。

设

$$T=\begin{bmatrix} \alpha_1 & \beta_2 & & & 0 \\ \beta_2 & \alpha_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \alpha_{n-1} & \beta_n \\ 0 & & & \beta_n & \alpha_n \end{bmatrix}$$

是问题2的解，并设 $\tilde{T}$ 是将 α_1 换成 $\tilde{\alpha}_1$ 之后得到的。则有

$$\tilde{\alpha}_1-\alpha_1=\mathrm{tr}(\tilde{T})-\mathrm{tr}(T)=\sum_{i=1}^{n}(\mu_i-\lambda_i). \qquad (3.1)$$

现在记 T 和 $\tilde{T}$ 的特征多项式分别为 $p_n(\lambda)$ 和 $q_n(\lambda)$，并记 T 之右下角的 $n-1$ 阶和 $n-2$ 阶主子阵的特征多项式分别为 $p_{n-1}(\lambda)$ 和 $p_{n-2}(\lambda)$，则由三对角矩阵的基本性质有

$$p_n(\lambda)=(\lambda-\alpha_1)p_{n-1}(\lambda)-\beta_2^2p_{n-2}(\lambda),$$

$$q_n(\lambda)=(\lambda-\tilde{\alpha}_1)p_{n-1}(\lambda)-\beta_2^2p_{n-2}(\lambda).$$

于是，有

$$p_n(\lambda)-q_n(\lambda)=(\tilde{\alpha}_1-\alpha_1)p_{n-1}(\lambda). \qquad (3.2)$$

将(3.1)代入(3.2)，并注意到

$$p_n(\lambda)=\prod_{i=1}^{n}(\lambda-\lambda_i),\quad q_n(\lambda)=\prod_{i=1}^{n}(\lambda-\mu_i),$$

就有

$$p_{n-1}(\lambda)\sum_{i=1}^{n}(\mu_i-\lambda_i)=\prod_{i=1}^{n}(\lambda-\lambda_i)-\prod_{i=1}^{n}(\lambda-\mu_i).\tag{3.3}$$

反过来，从给定的数据 λ_i 和 μ_i，由(3.3)可唯一地确定 $p_{n-1}(\lambda)$，而且易证 $p_{n-1}(\lambda)$ 的零点严格地分隔 $p_n(\lambda)$ 的 n 个零点(因 $p_{n-1}(\lambda_j)$ 的符号为 $(-1)^{n-j}$)。这样由定理1.2知，存在唯一的 Jacobi 矩阵 T，使得 T 的特征多项式是 $p_n(\lambda)$，而右下角的 $n-1$ 阶主子阵的特征多项式正好是由(3.3)确定的 p_{n-1}。而将如此得到的 T 之 a_1 换作

$$\tilde{a}_1=a_1+\sum_{i=1}^{n}(\mu_i-\lambda_i)\tag{3.4}$$

所得到的矩阵 $\tilde{T}$ 之特征多项式正好是 $q_n(\lambda)=\prod_{i=1}^{n}(\lambda-\mu_i)$。这也就证明了问题 2 之解存在唯一，而且可按如下步骤求得问题 2 的解：

(1) 计算

$$\omega_i=[p_{n-1}(\lambda_j)/p_n'(\lambda_j)]^{1/2},\quad j=1,2,\cdots,n,$$

其中 $p_{n-1}(\lambda)$ 由(3.3)定义，$p_n=\prod_{j=1}^{n}(\lambda-\lambda_i)$。

(2) 对 $\lambda_1,\cdots,\lambda_n$ 和 $q=(\omega_1,\cdots,\omega_n)^T$ 用上节的正交约化法求出 Jacobi 矩阵 T。

3.2 广对称 Jacobi 矩阵的特征值反问题

问题 3 给定 n 个实数

$$\lambda_1<\lambda_2<\cdots<\lambda_n,$$

求一个 n 阶广对称 Jacobi 矩阵 T，使得 T 的特征值就是给定的数 $\lambda_1,\cdots,\lambda_n$。

为了解决这一问题，我们先来讨论一下广对称 Jacobi 矩阵的基本性质。设 Jacobi 矩阵

$$T=\begin{bmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \ddots & & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & & \ddots & \alpha_{n-1} & \beta_n \\ & & & \beta_n & \alpha_n \end{bmatrix}$$

是广对称的，即

$$\begin{cases} \alpha_i = \alpha_{n-i+1}, & i=1,2,\cdots,[n/2], \\ \beta_i = \beta_{n-i+2}, & i=2,3,\cdots,[n/2]. \end{cases} \tag{3.5}$$

令 $k=[n/2]$。记 $S=E_k S^p E_k$，其中

$$S^p=\begin{bmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \alpha_{k-1} & \beta_k \\ & & & \beta_k & \alpha_k \end{bmatrix},$$

$E_k=[e_k,\cdots,e_1]\in\mathbb{R}^{k\times k}$ 为 k 阶反序单位矩阵。则由(3.5)可知 T 可写成如下形式：

(1) 当 $n=2k$ 时，

$$T=\begin{bmatrix} S^p & \beta_{k+1}e_k e_1^{\mathrm{T}} \\ \beta_{k+1}e_1 e_k^{\mathrm{T}} & S \end{bmatrix}\begin{matrix} k \\ k \end{matrix}; \tag{3.6}$$
$$\qquad\quad k \qquad\qquad\quad k$$

(2) 当 $n=2k+1$ 时，

$$T=\begin{bmatrix} S^p & \beta e_k & 0 \\ \beta e_k^{\mathrm{T}} & \alpha_{k+1} & \beta e_1^{\mathrm{T}} \\ 0 & \beta e_1 & S \end{bmatrix}\begin{matrix} k \\ 1 \\ k \end{matrix}, \tag{3.7}$$
$$\quad k \qquad 1 \qquad k$$

其中 $\beta=\beta_{k+1}=\beta_{k+2}$.

于是有：

(1) 当 $n=2k$ 时，

$$T=\frac{1}{2}\begin{bmatrix}E_k & -E_k\\ I_k & I_k\end{bmatrix}\begin{bmatrix}S+\beta e_1e_1^{\mathrm{T}} & 0\\ 0 & S-\beta e_1e_1^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}E_k & I_k\\ -E_k & I_k\end{bmatrix},\tag{3.8}$$

其中 $\beta=\beta_{k+1}$；

(2) 当 $n=2k+1$ 时

$$T=\frac{1}{2}\begin{bmatrix}0 & E_k & -E_k\\ \sqrt{2} & 0 & 0\\ 0 & I_k & I_k\end{bmatrix}\begin{bmatrix}\alpha_{k+1} & \sqrt{2}\beta e_1^{\mathrm{T}} & 0\\ \sqrt{2}\beta e_1 & S & 0\\ 0 & 0 & S\end{bmatrix}\times\begin{bmatrix}0 & \sqrt{2} & 0\\ E_k & 0 & I_k\\ -E_k & 0 & I_k\end{bmatrix},\tag{3.9}$$

其中 $\beta=\beta_{k+1}=\beta_{k+2}$。

这表明广对称 Jacobi 矩阵的特征值问题可以化为两个阶数等于原矩阵阶数一半的 Jacobi 矩阵的特征值问题。

由于 T 的次对角元素均为正数，故它的特征值互不相同，假设其为

$$\lambda_1<\lambda_2<\cdots<\lambda_n.$$

当 $n=2k+1$ 时，由(3.9)知，

$$\lambda(T)=\lambda(T_{k+1})\cup\lambda(S),$$

其中

$$T_{k+1}=\begin{bmatrix}\alpha_{k+1} & \sqrt{2}\beta e_1^{\mathrm{T}}\\ \sqrt{2}\beta e_1 & S\end{bmatrix}.$$

现在假定 T_{k+1} 和 S 的特征值分别为

$$\mu_1<\mu_2<\cdots<\mu_{k+1}\text{和}\nu_1<\nu_2<\cdots<\nu_k,$$

则由 S 是 T_{k+1} 的一个 k 阶主子阵，故有

$$\mu_1<\nu_1<\mu_2<\nu_2<\cdots<\nu_k<\mu_{k+1}.$$

从而有

$$\begin{cases}\mu_i = \lambda_{2i-1}, & i = 1,2,\cdots,k,k+1,\\ \nu_i = \lambda_{2i}, & i = 1,2,\cdots,k.\end{cases} \tag{3.10}$$

当$n = 2k$时，由(3.8)知

$$\lambda(T) = \lambda(S + \beta e_1 e_1^{\mathrm{T}}) \cup \lambda(S - \beta e_1 e_1^{\mathrm{T}}). \tag{3.11}$$

为了给出T的特征值与$S + \beta e_1 e_1^{\mathrm{T}}$和$S - \beta e_1 e_1^{\mathrm{T}}$的特征值之间的进一步关系，先证一个引理。

引理3.1 设A是$n \times n$实对称矩阵，它的特征值是$\lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_n$，再设$\beta$是一正数，$A - \beta e_1 e_1^{\mathrm{T}}$的特征值是$\tilde{\lambda}_1 \leqslant \tilde{\lambda}_2 \leqslant \cdots < \tilde{\lambda}_n$，则有

$$\tilde{\lambda}_1 \leqslant \lambda_1 \leqslant \tilde{\lambda}_2 \leqslant \cdots \leqslant \tilde{\lambda}_n \leqslant \lambda_n.$$

证明 对任意的$x = (\xi_1,\cdots,\xi_n)^{\mathrm{T}} \in \mathbb{R}^n$，有

$$x^{\mathrm{T}}(A - \beta e_1 e_1^{\mathrm{T}})x = x^{\mathrm{T}} A x - \beta \xi_1^2 \leqslant x^{\mathrm{T}} A x,$$

因而由 Hermite 矩阵的极小极大定理，有

$$\tilde{\lambda}_i \leqslant \lambda_i, \quad i = 1,2,\cdots,n. \tag{3.12}$$

设A的右下角的$n-1$阶主子阵的特征值是$\mu_1 \leqslant \cdots \leqslant \mu_{n-1}$，则有

$$\lambda_1 \leqslant \mu_1 \leqslant \cdots \leqslant \mu_{n-1} \leqslant \lambda_n$$

和

$$\tilde{\lambda}_1 \leqslant \mu_1 \leqslant \cdots \leqslant \mu_{n-1} \leqslant \tilde{\lambda}_n.$$

从而有

$$\tilde{\lambda}_i \geqslant \mu_{i-1} \geqslant \lambda_{i-1}, \quad i = 2,\cdots,n. \tag{3.13}$$

由(3.12)和(3.13)便知引理3.1的结论成立。

注3.1 在引理3.1中将β换作负数时，完全类似可证

$$\lambda_1 \leqslant \tilde{\lambda}_1 \leqslant \cdots \leqslant \lambda_n \leqslant \tilde{\lambda}_n.$$

设$S + \beta e_1 e_1^{\mathrm{T}}$和$S - \beta e_1 e_1^{\mathrm{T}}$的特征值分别为

$$\mu_1 < \mu_2 < \cdots < \mu_k \text{ 和 } \nu_1 < \nu_2 < \cdots < \nu_k.$$

则注意到$\beta = \beta_{k+1} > 0$，从引理3.1可推知

$$\nu_1 < \mu_1 < \nu_2 < \mu_2 < \cdots < \nu_k < \mu_k. \tag{3.14}$$

从(3.11)和(3.14)即知

$$\mu_i = \lambda_{2i}, \quad \nu_i = \lambda_{2i-1}, \quad i = 1, 2, \cdots, k. \tag{3.15}$$

此外，还有

$$\begin{aligned} 2\beta &= \mathrm{tr}(S + \beta e_1 e_1^{\mathrm{T}}) - \mathrm{tr}(S - \beta e_1 e_1^{\mathrm{T}}) \\ &= \sum_{i=1}^{k} (\lambda_{2i} - \lambda_{2i-1}). \end{aligned} \tag{3.16}$$

现在我们利用上述关于广对称 Jacobi 矩阵的基本性质来解决问题 3 .

当 $n = 2k + 1$ 时，由(3.10),(3.9)和(3.7)知，此时问题 3 等价于：求一个$k + 1$阶 Jacobi 矩阵

$$T_{k+1} = \begin{bmatrix} \alpha_{k+1} & \sqrt{2}\beta e_1^{\mathrm{T}} \\ \sqrt{2}\beta e_1 & S \end{bmatrix},$$

使得T_{k+1}的特征值是$\lambda_1, \lambda_3, \cdots, \lambda_{2k-1}, \lambda_{2k+1}$，而$S$的特征值是 $\lambda_2, \lambda_4, \cdots, \lambda_{2k}$. 这正好是我们曾讨论过的基本问题，因此，问题3有且仅有唯一的解，而且可用 § 2中所介绍的方法求解.

当$n = 2k$时，从(3.6),(3.8)和(3.15)知，问题 3 就等价于：求一个 k 阶 Jacobi 矩阵

$$S - \beta e_1 e_1^{\mathrm{T}},$$

使得它的特征值是 $\lambda_1, \lambda_3, \cdots, \lambda_{2k-1}$， 而且 $S + \beta e_1 e_1^{\mathrm{T}}$ 的特征值是 $\lambda_2, \cdots, \lambda_{2k}$. 这正好是我们曾讨论过的秩 1 修改问题，因此，此时问题 3 亦有唯一的解，而且亦可用 § 2中所介绍的方法来求解.

3.3 对角矩阵与秩1矩阵之和的特征值

问题 4 设 $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$, $\lambda_i \in \mathbb{R}$, $d \in \mathbb{R}^n$, $d \neq 0$. 求矩阵$\Lambda + dd^{\mathrm{T}}$ 的特征值.

当然这一问题可以应用对称 QR 方法来求解. 而这样做的第一步就是将 $\Lambda + dd^{\mathrm{T}}$ 约化为对称三对角矩阵. 如果按 Householder 方法来约化，就需运算量为 $O(n^3)$；但如果利用 $\Lambda + dd^{\mathrm{T}}$ 的特殊

性，应用上节所介绍的方法来约化，则只需运算量$O(n^2)$就可完成约化其为三对角矩阵的任务。

由上节的讨论知，对于对角加边矩阵

$$A=\begin{bmatrix} \alpha_0 & d^{\mathrm{T}} \\ d & \Lambda \end{bmatrix} \tag{3.17}$$

可以用正交约化法以$O(n^2)$次运算就可将A三对角化，即可求得n阶正交矩阵Q，使得

$$\begin{bmatrix} 1 & 0 \\ 0 & Q^{\mathrm{T}} \end{bmatrix} A \begin{bmatrix} 1 & 0 \\ 0 & Q \end{bmatrix} = \begin{bmatrix} \alpha_0 & \beta_1 e_1^{\mathrm{T}} \\ \beta_1 e_1 & T \end{bmatrix}. \tag{3.18}$$

比较(3.18)的两边可得

$$Q^{\mathrm{T}}\Lambda Q=T, \qquad \beta_1 Q e_1 = d, \qquad \beta_1 = \pm\|d\|_2.$$

由此可得

$$Q^{\mathrm{T}}(\Lambda+dd^{\mathrm{T}})Q=T+\beta_1^2 e_1 e_1^{\mathrm{T}},$$

即这样得到的Q亦将$\Lambda+dd^{\mathrm{T}}$约化为对称三对角矩阵。因此，可按如下算法求$\Lambda+dd^{\mathrm{T}}$的特征值。

算法3.1

(1) 用“驱逐出境法”或“Rutishauser方法”将(3.17)约化为(3.18)的形式，求得对称三对角矩阵T和常数β。

(2) 用带Wilkinson位移的对称QR方法求矩阵$T+\beta^2 e_1 e_1^{\mathrm{T}}$的特征值。

习 题

1. 设T是实对称三对角矩阵，$\tilde{T}$是将T的次对角元素取绝对值之后得到的矩阵。证明T与$\tilde{T}$有相同的特征值。

2. 设A为对角加边矩阵

$$A=\begin{bmatrix} \alpha & b^{\mathrm{T}} \\ b & M \end{bmatrix},$$

其中$b=(\beta_2,\cdots,\beta_n)^{\mathrm{T}}$，$M=\mathrm{diag}(\mu_1,\cdots,\mu_{n-1})$，$\mu_1<\mu_2<\cdots<\mu_{n-1}$.

试证实数 $\lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_n$ 是 A 的特征值的充分必要条件是：

(1) $\lambda_1 \leqslant \mu_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \mu_{n-1} \leqslant \mu_n$；

(2) $$\beta_{i+1}^2 = -\prod_{j=1}^{n}(\mu_i - \lambda_j) \Big/ \prod_{\substack{j=1 \\ j \neq i}}^{n-1}(\mu_i - \mu_j),$$
$$i = 1, 2, \cdots, n-1.$$

3. 设 $\mathscr{P}_N = \operatorname{span}\{1, x, \cdots, x^{N-1}\}$，即 $\mathscr{P}_N$ 表示全体次数不超过 $N-1$ 的实系数多项式的全体，$\xi_1, \cdots, \xi_N$ 是 N 个互不相同的实数，$\omega_1, \cdots, \omega_N$ 是 N 个正数。定义

$$\langle p, q \rangle = \sum_{i=1}^{N} \omega_i p(\xi_i) q(\xi_i), \quad p, q \in \mathscr{P}_N.$$

证明 $\langle \cdot, \cdot \rangle$ 是线性空间 $\mathscr{P}_N$ 上的一个内积。

4. 证明存在唯一的首一多项式序列 $\{q_i(x)\}_0^N$ 满足：

(1) $q_i(x)$ 的次数是 i；

(2) $\langle q_i, q_j \rangle = 0$，$i \neq j$。

其中 $\langle \cdot, \cdot \rangle$ 是习题 3 所定义的内积。

5. 设 $\{q_i(x)\}_0^N$ 是习题 4 所述的首一多项式列，证明它们满足如下的三项递推公式：

$$q_0(x) \equiv 1, \qquad q_1(x) = x - \alpha_1,$$
$$q_i(x) = (x - \alpha_i) q_{i-1}(x) - \beta_i^2 q_{i-2}(x), \qquad i = 2, 3, \cdots, N,$$

其中

$$\alpha_i = \langle q_{i-1}, x q_{i-1} \rangle / \langle q_{i-1}, q_{i-1} \rangle, \qquad i = 1, 2, \cdots, N,$$
$$\beta_i = [\langle q_{i-1}, q_{i-1} \rangle / \langle q_{i-2}, q_{i-2} \rangle]^{\frac{1}{2}}, \qquad i = 2, \cdots, N,$$

$\langle \cdot, \cdot \rangle$ 如习题 3 所定义。

6. 设 $p_N(x)$ 和 $p_{N-1}(x)$ 是两个次数分别为 N 和 $N-1$ 的首一多项式，证明 p_N 和 p_{N-1} 是习题 4 所述的首一多项式序列中两个次数最高的多项式的充分必要条件是：

(1) ξ_i 是 $p_N(\lambda)$ 的根，即 $p_N(\xi_i) = 0$，$i = 1, \cdots, N$；

(2) $p'_N(\xi_i)p_{N-1}(\xi_i)>0$，且存在正数 γ 使得

$$\omega_i=\gamma/p'_N(\xi_i)p_{N-1}(\xi_i),\quad i=1,2,\cdots,N.$$

7. 利用习题6的结论证明问题1有且仅有一个解，并据此给出一种求解问题1的算法.

8. 给定实数列 $\{\lambda_i^{(k)}\}_{i=1}^{k}(k=n-r,\cdots,n)$，满足

$$\lambda_i^{(k)}\leqslant\lambda_i^{(k-1)}\leqslant\lambda_{i+1}^{(k)},\quad i=1,2,\cdots,k-1,\quad k=n-r+1,\cdots,n.$$

试设计一种算法，构造一个带宽为 $2r+1$ 的实对称矩阵 A，使得 A 的 k 阶顺序主子阵 A_k 的特征值是 $\lambda_1^{(k)},\cdots,\lambda_k^{(k)}(k=n-r,\cdots,n)$.

9. 给定一个正数 β 和 $2n-1$ 个实数

$$\lambda_1<\mu_1<\lambda_2<\cdots<\mu_{n-1}<\lambda_n.$$

试给出一种算法，计算一个周期Jacobi矩阵

$$J_n=\begin{bmatrix}\alpha_1 & \beta_2 & & & \beta_1\\ \beta_2 & \alpha_2 & \ddots & & 0\\ & \ddots & \ddots & \ddots & \\ 0 & & \ddots & \alpha_{n-1} & \beta_n\\ \beta_1 & & & \beta_n & \alpha_n\end{bmatrix},\quad \beta_i>0,$$

使得 J_n 的特征值是 $\lambda_1,\cdots,\lambda_n$，$J_n$ 的 $n-1$ 阶顺序主子阵的特征值是 $\mu_1,\cdots,\mu_{n-1}$，并且有

$$\beta_1\beta_2\cdots\beta_n=\beta.$$

参考文献

[1] 蔡大用，数值代数，清华大学出版社，1987.

[2] 曹志浩，矩阵特征值问题，上海科学技术出版社，1980.

[3] 曹志浩、张玉德、李瑞遐，矩阵计算和方程求根，高等教育出版社，1979.

[4] 何旭初、孙文瑜，广义逆矩阵引论，江苏科学技术出版社，1991.

[5] 黄友谦，数值试验，高等教育出版社，1989.

[6] 胡家赣，线性代数方程组的迭代解法，科学出版社，1991.

[7] 蒋尔雄、高坤敏、吴景琨，线性代数，人民教育出版社，1978.

[8] 蒋尔雄，对称矩阵计算，上海科学技术出版社，1984.

[9] 沈启钧，把对称镶边对角阵约化成三对角阵的一个新的快速算法，1991年计算数学天津会议文集，521—523.

[10] 孙继广，矩阵扰动分析，科学出版社，1987.

[11] 王德人、杨忠华，数值逼近引论，高等教育出版社，1990.

[12] 王则可，同伦算法的几何理论，高等学校计算数学学报，第3期，1988，202—210.

[13] 王则可、高庆堂，同伦方法引论，重庆出版社，1990.

[14] 徐树方，关于SSOR迭代法应用于最小二乘问题时的收敛定理的一个注记，高等学校计算数学学报，第1期，1993，95—98.

[15] J. O. Aasen, On the reduction of a symmetric matrix to tridiagonal form, *BIT*, 11(1971), 233—242.

[16] O. Axelsson and V. A. Barker, *Finite Element Solutions of Boundary Value Problems: Theory and Computation*, Academic Press, New York and London, 1984.

[17] O. Axelsson and G. Lindskog, On the rate of convergence of the preconditioned conjugate gradient method, *Numer. Math.*, 48(1986), 499—523.

[18] S. Barnett, *Matrices: Methods and Applications*, Clarenden Press, Oxford, 1990.

[19] Å. Björck and V. Pereyra, Solution of Vandermonde systems of equations, *Math. Comp.*, 24(1970), 893—903.

[20] Å. Björck, *Least Squares Methods*, in: P. G. Ciarlet and

J. L. Lions eds., Handbook of Numerical Analysis, Vol.1, North-Holand, Amsterdam, 1990.

[21] D. L. Boley and G. H. Golub, A survey of matrix inverse eigenvalue problems, *Inverse Problems*, 3(1987), 595—622.

[22] P. N. Brown, A theoretical comparison of the Arnoldi and GMRES algorithms, *SIAM J. Sci. Stat. Comput.*, 12(1991), 58—78.

[23] T. F. Chan, Rank revealing QR-factorizations, *Lin. Alg. and Its Applic.*, 88/89 (1987), 67—82.

[24] M. T. Chu, A simple application of the homotopy method to symmetric eigenvalue problems, *Lin. Alg. and Its Applic.*, 59(1984), 85—90.

[25] M. T. Chu, A note on the homotopy method for linear algebraic eigenvalue problems, *Lin. Alg. and Its Applic.*, 105(1988), 225—236.

[26] A. K. Cline, C. B. Moler, G. W. Stewart and J. H. Wilkinson, An estimate for the condition number of a matrix, *SIAM J. Numer. Anal.*, 16(1979), 368—375.

[27] P. Concus and G. H. Golub, *A generalized conjugate gradient method for nonsymmetric systems of linear equations*, in Lecture Notes in Economics and Mathematical Systems 134, R. Glowinski and J. L. Lions eds., Springer-Verlag, Berlin (1976), 56—65.

[28] P. Concus, G. H. Golub and G. Meurant, Block preconditioning for conjugate gradient method, *SIAM J. Sci. Stat. Comput.*, 6(1985), 220—252.

[29] J. K. Cullum and R. A. Willoughby, *Lanczos Algorithms for Large Symmetric Eigenvalue Computations*, Vol.1 Theorem and Vol.2 Programs, Birkhaüser, Boston, 1985.

[30] J. J. M. Cuppen, A divide and conquer method for the symmetric eigenproblem, *Numer. Math.*, 36(1981), 177—195.

[31] J. J. Dongarra and D. C. Sorensen, A fully parallel algorithm for symmetric eigenvalue problem, *SIAM J. Sci. and Stat. Comp.*, 8(1987), S139—S154.

[32] S. C. Eisenstat, H. C. Elman and M. H. Schultz, Varia-

tional iterative methods for nonsymmetric systems of linear equations, *SIAM J. Numer. Anal.*, 20(1983), 345—357.

[33] S. C. Eisenstat, A note on the generalized conjugate gradient method, *SIAM J. Numer. Anal.*, 20(1983), 358—361.

[34] P. E. Gill, W. Munay and M. H. Wright, *Numerical Linear Algebra and Optimization*, Addison-Wesley Publishing Company, 1991.

[35] G. H. Golub and D. P. O'Leary, Some history of the conjugate gradient and Lanczos algorithms: 1948—1976, *SIAM Review*, 31(1989), 50—102.

[36] G. H. Golub and C. F. Van Loan, An analysis of the total least squares problems, *SIAM J. Numer. Anal.*, 17(1980), 883—893.

[37] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Second edition, The Johns Hopkins University Press, Baltimore and London, 1989.

[38] L. A. Hagemen and D. M. Young, *Applied Iterative Methods*, Academic Press, New York and London, 1981.

[39] O. H. Hald, Inverse eigenvalue problems for Jacobi matrices, *Lin. Alg. and Its Applic.*, 14(1976), 63—85.

[40] M. R. Hestenes and E. Stiefel, Methods of conjugate gradients for solving linear systems, *J. Res. Nat. Bur. Standards*, 49(1952), 409—436.

[41] M. R. Hestenes, *Conjugate Direction Methods in Optimization*, Springer-Verlag, 1980.

[42] H. Hochstadt, On some inverse problems in matrix theory, *Arch. Math.*, 18(1967), 201—207.

[43] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, London, 1985.

[44] O. G. Johnson, C. A. Micchelli and G. Paul, Polynomial preconditions for conjugate gradient calculations, *SIAM J. Numer. Anal.*, 20(1983), 362—376.

[45] W. Kahan, Numerical linear algebra, *Canad. Math. Bull.*, 9(1966), 757—801.

[46] S.Kaniel, Estimates for some computational techniques

in linear algebra, *Math.* Comp., 20(1966), 369—378.

[47] D.S.Kershaw, The incomplete Cholesky-conjugate gradient method for the iterative solution of systems of linear equ-ations, *J.Comput.Physics*, 26(1978), 43—65.

[48] T.Y. Li and N. Rhee, Homotopy algorithm for symmetric eigenvalue problems, *Numer.Math.*, 55(1989), 256—280.

[49] T.A.Manteuffel, The Tschebychev iteration for nonsymme-tric linear systems, *Numer.Math.*, 28(1977), 307—327.

[50] G.Markham,Conjugate gradient type methods for indefinite, asymmetric, and complex systems, *SIMA J.Numer. Anal.*, 10(1990), 155—170.

[51] J.A.Meijerink and H.A.Van der Vorst,An iterative solution for linear systems of which the coefficient matrix is a sym-metric M-matrix, *Math.Comput.*, 31(1977), 148—162.

[52] W. Niethammer, J.de Phillis and R.S.Varga, Convergence of block iterative methods applied to sparse least squares problems, *Lin.Alg.and Its Applic.*, 58(1984), 327—341.

[53] C.C.Paige, *The computation of eigenvalues and eigenvectors of very large sparse matrices*, Ph.D.thesis,London University, London, England, 1971.

[54] C.C.Paige and M.A.Saunders, Solution of sparse indefinite systems of linear equtions, *SIAM J.Numer. Anal.*, 12(1975), 617—629.

[55] C.C.Paige, Error analysis of the Lanczos algorithm for tri-diagonalizing symmetric matrix, *J.Inst. Math. Applic.*, 18 (1976), 341—349.

[56] C.C. Paige, Accuracy and effectiveness of the Lanczos al-gorithm for the symmetric eigenproblem, *Lin. Alg. and Its Applic.*, 34(1980), 235—258.

[57] B.N.Parlett,*The Symmetric Eigenvalue Problem*,Prentic-Hall, Englewood Cliffs, NJ, 1980.

[58] D.J.Rose and R.A. Willoughby eds., *Sparse Matrices and Their Applications*, Plenum Press, New York and London, 1972.

[59] Y.Saad,On the rates of convergence of the Lanczos and the

block Lanczos methods, *SIAM J.Numer. Anal.*, 17(1980), 689—706.

[60] Y.Saad and M.H. Schultz, GMRES: A generalized minimal resdual algorithm for solving nonsymmetric linear systems, *SIAM J.Sci.Stat.Comput.*, 7(1986), 856—869.

[61] G.W.Stewart,On the continuity of the generalized inverse, *SIAM J.Appl.Math.*, 17(1969), 33—45.

[62] G.W.Stewart, *Introduction to Matrix Computations*,Academic Press, New York and London, 1973.

[63] G.W.Stewart,The efficient generation of random orthogonal matrices with applications to condition estimators, *SIAM J.Numer. Anal.*, 17(1980), 403—409.

[64] G.W. Stewart and J. G. Sun, *Matrix Perturbation Theory*, Academic Press, New York and London, 1990.

[65] J.G. Sun, On the condition number of a nondefective multiple eigenvalue, *Numer.Math.*, 61(1992), 265—275.

[66] A.Van der Sluis and H.A. Van der Vorst, The rate of convergence of cojugate gradients, *Numer. Math.*, 48(1986), 543—560.

[67] H.A.Van der Vorst and K.Dekker, Conjugate gradient type methods and preconditioning,*J.Comput.Appl.Math.*, 24(1988), 73—87.

[68] J.M. Varah, On the numerical solutions of ill-conditioned linear systems with applications to ill-posed problems, *SIAM J.Numer.Anal.*, 10(1973), 549—565.

[69] R.S.Varga, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[70] H.F. Walker, Implementation of the GMRES method using Householder trasformations, *SIAM J. Sci. Stat. Comput.* 9 (1988), 152—163.

[71] J.H.Wilkinson, *The Algebraic Eigenvalue Problem*,Clarendon Press, Oxford, 1965.

[72] S.F. Xu, A stability analysis of the Jacobi matrix inverse eigenvalue problem, *BIT*, 33(1993), 695—702.

[73] D.M.Young, *Iterative Solution of Large Linear Systems*, Academic Press, New York and London, 1971.

索　引